

# PATENT ABSTRACTS OF JAPAN

(11)Publication number : **07-129331**

(43)Date of publication of application : **19.05.1995**

(51)Int.Cl.

**G06F 3/06**

**G06F 3/06**

**G11B 19/02**

(21)Application number : **05-276317**

(71)Applicant : **FUJITSU LTD**

(22)Date of filing : **05.11.1993**

(72)Inventor : **MATOKA TATSUO**

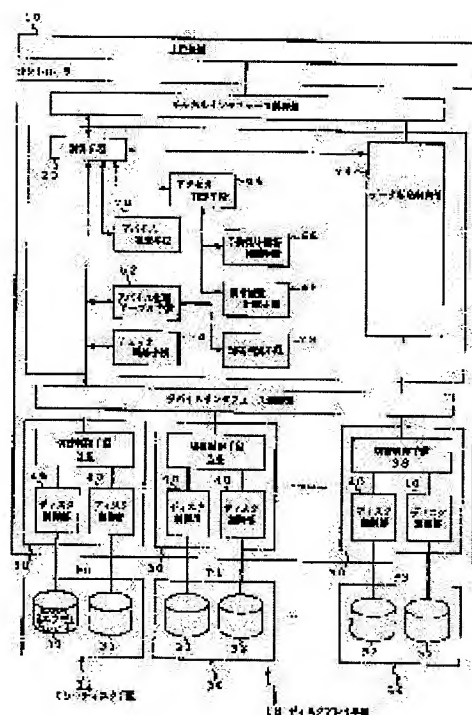
## (54) DISK ARRAY DEVICE

### (57)Abstract:

**PURPOSE:** To instantaneously process a processing request without losing data even if plural disk devices operated in parallel are simultaneously failed.

**CONSTITUTION:** A mirror disk device 36 consisting of a pair of disk devices 32 is used as one of constitutional elements for a disk array 18. A current device and stand-by devices are allocated to mirror disk devices 32, data are written in both the current and stand-by devices and data are read out only from the current device. When the generation of a fault in the current device is judged, allocation is previously switched from the current device to a stand-by device.

Simulation for inspecting the disk array 18 is executed in an idle state to collect fault information. In the mirror disk constitution, RAID for storing parities can also be executed. When mirror disk constitution is not used for a current device, data are preparatorily copied at the time of judging the generation of a fault in the current device to dynamically obtain mirror disk constitution.



## LEGAL STATUS

[Date of request for examination] 05.09.1997

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3078972

[Date of registration] 16.06.2000

[Number of appeal against examiner's decision of rejection]

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号

特開平7-129331

(43)公開日 平成7年(1995)5月19日

(51)Int.Cl. <sup>6</sup>	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	5 4 0	7165-5B		
	3 0 6 B			
G 1 1 B 19/02	5 0 1 F	7525-5D		

審査請求 未請求 請求項の数23 O L (全 30 頁)

(21)出願番号 特願平5-276317

(22)出願日 平成5年(1993)11月5日

(71)出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中1015番地

(72)発明者 的場 辰夫

神奈川県川崎市中原区上小田中1015番地

富士通株式会社内

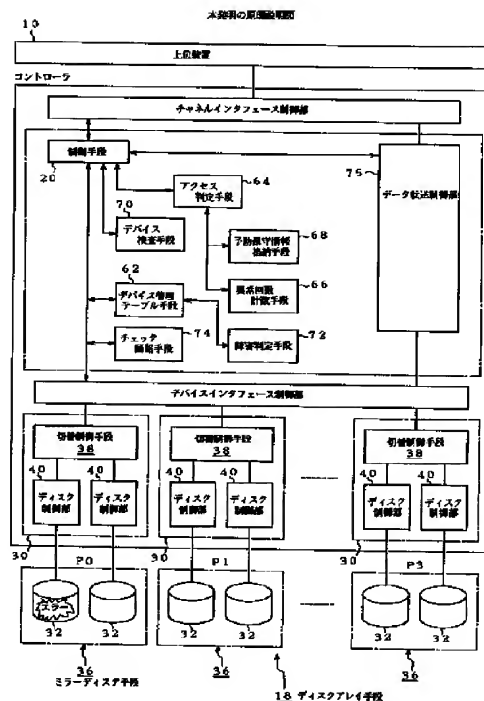
(74)代理人 弁理士 竹内 進 (外1名)

(54)【発明の名称】 ディスクアレイ装置

(57)【要約】

【目的】並列動作しているディスク装置の複数が同時に故障してもデータを喪失することなく瞬時に処理要求に対処可能とする。

【構成】ディスクアレイ18の構成要素としてディスク装置32を2台で一組としたミラーディスク装置36を用いる。ミラーディスク装置には現用と予備用が割当られ、データ書込みは現用と予備用の両方に行い、データ読出しは現用から行う。ディスク装置の障害発生を判定し、事前に現用から予備へ割当てを切替えておく。アイドル状態でディスクアレイを検査するシミュレーションを行って障害情報を収集する。ミラーディスク構成でパリティを格納するRAID3又は5としてもよい。更に現用はミラーディスクとせず、現用の障害判定時に予備にデータを複写し動的にミラーディスク構成とする。



## 【特許請求の範囲】

【請求項1】同一データを格納する2台のディスク装置(36)を備えたミラーディスク手段(36)を並列アクセス可能に複数配置したディスクアレイ手段(18)と、上位装置(10)からの書込データをストライピングした後に前記ディスクアレイ手段(36)に並列書込みすると共に、前記ディスクアレイ手段(36)からの並列読出データを合成して前記上位装置(10)に転送する制御手段(20)と、を備えたことを特徴とするディスクアレイ装置。

【請求項2】請求項1記載のディスクアレイ装置に於いて、更に、前記ディスクアレイ手段(18)に設けた複数のディスク装置の状態を管理するデバイス管理テーブル手段(62)を設け、前記制御手段(20)は、上位装置(10)からのアクセス要求時に前記デバイス管理テーブル手段(62)を参照して前記ディスクアレイ手段(18)に対する処理を実行することを特徴とするディスクアレイ装置。

【請求項3】請求項1記載のディスクアレイ装置に於いて、前記デバイス管理テーブル手段(62)に、前記ミラーディスク手段(36)を設けたディスク装置の一方を現用とし、他方を予備用とする情報を登録し、前記制御手段(20)は、リードアクセス時に前記デバイス管理テーブル手段(62)の参照で前記ミラーディスク手段(36)の現用側のディスク装置からデータを読み出し、ライトアクセス時には現用および予備用の両方のディスク装置にデータを書込むことを特徴とするディスクアレイ装置。

【請求項4】請求項3記載のディスクアレイ装置に於いて、前記ディスクアレイ手段(18)は、前記制御手段(20)からのリード動作の指示に対し前記ミラーディスク手段(36)の現用側のディスク装置に切替えてデータを読み出し、ライト動作の指示に対しては現用および予備用の両方のディスク装置にデータを書込むように切替える切替制御手段(40)を備えたことを特徴とするディスクアレイ装置。

【請求項5】請求項3記載のディスクアレイ装置に於いて、前記ディスクアレイ手段(18)は前記制御手段(20)に対し前記ミラーディスク手段(36)を構成する2台のディスク装置の各々を個別のポートを介して並列的に接続し、前記制御手段(20)はリード動作時には、前記ミラーディスク手段(36)の一方のポートによる現用側のディスク装置のアクセスでデータを読み出し、ライト動作時には、両方のポートによる現用および予備用の両方のディスク装置のアクセスでデータを書込むことと

徴とするディスクアレイ装置。

【請求項6】請求項5記載のディスクアレイ装置に於いて、前記制御手段(20)は、前記上位装置(10)からの書込要求時に、書込データを前記ミラーディスク手段(36)の並列アクセス台数に基づいてストライピングした後に各ストライピングデータと同一データを生成し、一対の同ストライピングデータを前記ミラーディスク手段(36)の2つのポートに転送することを特徴とするディスクアレイ装置。

【請求項7】請求項1記載のディスクアレイ装置に於いて、前記ディスクアレイ手段(18)は、着脱自在な少くとも2台のディスクモジュール(50)を内蔵したディスクユニット(46)と、前記ディスクユニット(46)の収納部を複数備え、予め定めたアレイ構成に従って複数台の前記ディスクユニット(46)を実装した装置筐体(44)と、で構成されたことを特徴とするディスクアレイ装置。

【請求項8】請求項7記載のディスクアレイ装置に於いて、前記ディスクユニット(46)は、着脱自在な前記ディスクモジュール(50)に対する共通部として電源部および冷却手段を備えたことを特徴とするディスクアレイ装置。

【請求項9】請求項7記載のディスクアレイ装置に於いて、前記ディスクモジュール(50)は、少なくともディスク媒体、ディスク回転機構、ヘッド機構を備えたことを特徴とするディスクアレイ装置。

【請求項10】請求項7記載のディスクアレイ装置に於いて、前記ディスクユニット(46)に、前記制御手段(20)からのリード動作の指示に対し前記ミラーディスク手段(36)の現用側のディスク装置に切替えてデータを読み出し、ライト動作の指示に対しては現用および予備用の両方のディスク装置にデータを書込むように切替える切替制御手段(40)の回路ユニットを内蔵したことを特徴とするディスクアレイ装置。

【請求項11】請求項7記載のディスクアレイ装置に於いて、前記ディスクユニット(46)を、媒体サイズの異なる単一のディスクモジュールを内蔵した他のディスクユニットの収納を予定したディスクアレイ筐体を実装してミラーディスク手段(36)を構成要素とするディスクアレイ構成を実現したことを特徴とするディスクアレイ装置。

【請求項12】請求項2記載のディスクアレイ装置に於いて、更に、前記ディスクアレイ手段(12)に対するアクセスの実行に対しチェック回路手段(74)において異常が発生

したか否かを判定するアクセス判定手段(64)と、前記アクセス判定手段(64)で判定した異常の回数を計数する異常回数数計数手段(66)と、前記異常回数計数手段(66)の異常回数が所定の閾値以上となったときにデバイス障害を判定し、前記デバイス管理テーブル手段(62)に障害情報を登録する障害判定手段(72)と、を設けたことを特徴とするディスクアレイ装置。

【請求項13】請求項12記載のディスクアレイ装置に於いて、更に、前記アクセス判定手段(64)の異常判定に対応して作成した予防保守情報を前記上位装置(10)で認識可能な状態に保持する予防保守情報格納手段(68)を設けたことを特徴とするディスクアレイ装置。

【請求項14】請求項2記載のディスクアレイ装置に於いて、更に、前記制御手段(20)の空き状態で、前記ディスクアレイ手段(18)に対し擬似的なアクセス動作を実行して各ディスク装置の状態を検査するデバイス検査手段(70)を設けたことを特徴とするディスクアレイ装置。

【請求項15】請求項14記載のディスクアレイ装置に於いて、前記デバイス検査手段(70)は、前記ディスクアレイ手段(18)の特定データ領域に対しダミーデータを用いた書込動作を実行することを特徴とするディスクアレイ装置。

【請求項16】請求項14記載のディスクアレイ装置に於いて、前記デバイス検査手段(70)は、前記ディスクアレイ手段(18)に対する擬似的なアクセス動作でアクセス判定手段(64)が異常を判定するごとに異常回数計数手段(66)に異常回数を計数させ、該異常回数が所定の閾値を越えて障害判定手段(72)で障害発生を判定した場合に、前記デバイス管理テーブル手段(62)に障害情報を登録することを特徴とするディスクアレイ装置。

【請求項17】請求項12記載のディスクアレイ装置に於いて、更に、前記デバイス検査手段(70)による擬似的なアクセス動作でアクセス判定手段(64)が判定した異常情報に対応した予防保守情報を前記上位装置(10)で認識可能な状態に保持する予防保守情報格納手段(68)を設けたことを特徴とするディスクアレイ装置。

【請求項18】請求項12又は16記載のディスクアレイ装置に於いて、前記制御手段(20)は、リード動作の実行時に前記デバイス管理テーブル手段(62)の障害情報を参照して障害ディスク装置を認識した場合、障害ディスク装置が現用側であれば予備用からのリード動作に切替え、障害ディスク装置が予備用であれば現用からのリード動作を

維持することを特徴とするディスクアレイ装置。

【請求項19】請求項12又は16記載のディスクアレイ装置に於いて、前記制御手段(20)は、ライト動作の実行時に前記デバイス管理テーブル手段(62)の障害情報を参照して障害ディスク装置を認識した場合、該障害ディスク装置をライト動作の対象から切り離すことを特徴とするディスクアレイ装置。

【請求項20】同一データを格納する2台のディスク装置(36)を備えたミラーディスク手段(36)を並列アクセス可能に複数配置したディスクアレイ手段(18)と、上位装置(10)からの書込データをストライピングしてパリティデータと共に前記ディスクアレイ手段(36)に並列に書込み、前記ディスクアレイ手段(36)からの並列読出データを合成して前記上位装置(10)に転送する制御手段(20)と、を備えたことを特徴とするディスクアレイ装置。

【請求項21】複数のディスク装置(32)を並列アクセス可能に配置したディスクアレイ手段(18)と、前記ディスクアレイ手段(18)の同一ランクに属する少なくとも1台のディスク装置を予備用に割り当てると共に他の複数ディスク装置をデータ格納用に割り当て、上位装置(10)からの書込データをストライピングし前記データ格納用ディスク装置に並列的に書込み、前記データ格納用ディスク装置からの並列読出データを合成して前記上位装置(10)に転送する制御手段(20)と、前記ディスクアレイ手段(12)に対するアクセスの実行に対しチェック回路手段(74)において異常が発生したか否かを判定するアクセス判定手段(64)と、前記アクセス判定手段(64)で判定した異常の回数を計数する異常回数数計数手段(66)と、前記異常回数計数手段(66)の異常回数が、通常の障害判定に用いる閾値より小さい所定の閾値以上となったとき、近い将来、障害発生の可能性があることを予測的に判定する障害判定手段(72)と、前記障害判定手段(72)の障害の予測判定時に前記予備用ディスク装置を選択して障害ディスク装置の格納データを複写し、該複写後に障害ディスク装置と前記予備用ディスク装置でミラーディスク手段(36)を構成し、前記制御手段(20)に両方のディスク装置への同時データの書込みといずれか一方のディスク装置からのデータ読出しをセットする構成制御手段と、を備えたことを特徴とするディスクアレイ装置。

【請求項22】請求項21のディスクアレイ装置に於いて、前記構成制御手段は、予備用ディスク装置を現用として前記制御手段(20)によるデータ読出しを行わせることを特徴とするディスクアレイ装置。

【請求項23】請求項21のディスクアレイ装置に於いて

て、  
前記障害判定手段(72)は、障害の予測判定が行われたことを外部に出力表示させる手段を備えたことを特徴とするディスクアレイ装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、アレイ状に構成された複数のディスク装置を上位装置からの命令に基づいて並列的にアクセスするディスクアレイ装置に関し、特にディスク障害が発生してもデータの復元が瞬時にできるディスクアレイ装置に関する。近年の計算機システムにおける高信頼性の要求に伴い、磁気ディスク装置を用いた入出力サブシステムでの高信頼性が要求されている。

【0002】このような高信頼性を実現するため、複数のディスク装置を並列的に動作させることでデータ及び冗長情報を格納するようにしたディスクアレイ装置が実用化されている。ディスクアレイ装置は、障害等により1台のディスクのデータが失われた場合、代替先として準備している予備のディスク装置に残りのディスク装置のデータを用いて失われたデータを修復することができ、高い信頼性が得られる。

【0003】しかし、同時にディスクアレイの中の複数のディスク装置で異常が発生した場合には、もはやデータの修復は不可能であり、この点の改善が望まれる。

【0004】

【従来の技術】近年、計算機システムの外部記憶装置として、記録の不揮発性、大容量性、データ転送の高速性等の特長を持つ磁気ディスク装置、光ディスク装置等のディスク装置が広く用いられている。ディスク装置に対する要求は、高速データ転送、信頼性重視、大容量性、低価格である。これらの要求を満たすものとして、ディスクアレイ装置が注目されてきている。

【0005】ディスクアレイ装置とは、小型ディスク装置を数台から数十台並べ、複数のディスク装置に分散してデータを記録して、並列的にアクセスする装置である。ディスクアレイ装置で並列的に複数のディスク装置にデータ転送を行えば、一台のディスク装置の場合と比べて、ディスクの台数倍の高速データ転送が可能になる。

【0006】また、データに加えて、パリティデータなどの冗長な情報を付け加えて記録しておくことで、ディスク装置の故障等を原因とするデータエラーの検出と訂正が可能となり、ディスク装置の内容を二重化して記録する方法と同程度の信頼性重視を、二重化より低価格で実現することができる。従来、カルフォルニア大学バークレイ校のデビット・A・パターソン(David A. Patterson)らは、高速に大量のデータを多くのディスクにアクセスし、ディスク故障時におけるデータの冗長性を実現するディスクアレイ装置について、レベル1からレベル5までに分類付けを行って評価した論文を発表してい

る(ACMSIGMOD Conference, Chicago, Illinois, June 1-3, 1988 P109-P116)。

【0007】このデビット・A・パターソンらが提案したディスクアレイ装置を分類するレベル1～5は、RAID (Redundant Arrays of Inexpensive Disks) 1～5と略称される。RAID 1～5を簡単に説明すると次のようになる。RAID 0は、データの冗長性をもたないディスクアレイ装置であり、デビット・A・パターソンらの分類には含まれていないが、これを仮にRAID 0と呼ぶ。

【0008】RAID 1は、2台のディスク装置を1組として同一データを書込むミラーディスク装置であり、ディスク装置の利用効率が低いながらも冗長性をもっており、簡単な制御で実現できるため、広く普及している。RAID 2は、データをビットやバイト単位でストライピング(分割)し、それぞれのディスク装置に並列に読み書きを行う。ストライピングしたデータは全てのディスク装置で物理的に同じセクタに記録する。

【0009】データ用ディスク装置の他にハミングコードを記録するためのディスク装置を持ち、ハミングコードから故障したディスク装置を特定して、データを復元する。しかし、実用化されていない。RAID 3は、データをビット又はバイト単位にストライピングしてパリティを計算し、ディスク装置に対しデータおよびパリティを並列的に書込む。

【0010】RAID 3は、大量のデータを連続して扱う場合には有効であるが、少量のデータをランダムにアクセスするランザクシオン処理のような場合には、データ転送の高速性が生かせず、効率が低下する。RAID 4は、1つのデータをセクタ単位にストライピングして同じディスク装置に書込む。

【0011】パリティは固定的に決めたディスク装置に格納している。データ書込みは、書込み前のデータとパリティを読み出してから新パリティを計算して書き込むため、1度の書込みについて、合計4回のアクセスが必要になる。また書込みの際に必ずパリティ用のディスク装置へのアクセスが起きるため、複数のディスク装置の書込みを同時に実行できない。

【0012】このようにRAID 4の定義は行われているが、メリットが少ないため現在のところ実用化の動きは少ない。RAID 5は、パリティ用のディスク装置を固定しないことで、並列的なリード、ライトを可能にしている。即ち、セクタごとにパリティの置かれるディスク装置が異なっている。パリティディスクが重複しなければ異なるディスク装置にセクタデータを並列的に書込むことができる。

【0013】このようにRAID 5は非同期に複数のディスク装置にアクセスしてリード又はライトを実行できるため、少量データをランダムにアクセスするランザクシオン処理に向いている。

【0014】

【発明が解決しようとする課題】ところで、現在、実用化が進められているRAID3およびRAID5の動作形態をもつディスクアレイ装置にあっては、例えば1つのパリティグループを構成するランクごとに予備用のディスク装置を準備しておき、1台のディスク装置が故障した場合には、他の正常なディスク装置のデータを読み出して故障ディスク装置のデータを予備用ディスク装置に修復するようにしている。

【0015】このためディスク装置の故障でデータが失われても復元できるため、極めて高い信頼性が得られる。しかしながら、同時に複数のディスク装置で故障が発生してデータが失われた場合には、残りの正常なディスク装置から失われたデータを修復することはできず、システムダウンとなってしまう問題があった。

【0016】本発明は、このような従来の問題点に鑑みてなされたもので、並列動作しているディスク装置の複数の同時に故障してもデータを喪失することなく瞬時に処理要求に対処できるディスクアレイ装置を提供することを目的とする。また本発明の他の目的は、ディスク装置のエラー状態を監視して障害発生の可能性が高くなった場合には事前に障害対処状態に切替えて障害発生を未然に防止するようにしたディスクアレイ装置を提供する。

【0017】

【課題を解決するための手段】図1は本発明の原理説明図である。本発明は、同一データを格納する2台のディスク装置36を備えたミラーディスク手段36を並列アクセス可能に複数配置したディスクアレイ手段18と、上位装置10からの書込データをストライピングした後にディスクアレイ手段18に並列書込みすると共に、ディスクアレイ手段18からの並列読出データを合成して前記上位装置10に転送する制御手段20とで構成される。

【0018】さらに、ディスクアレイ手段18に設けた複数のディスク装置の状態を管理するデバイス管理テーブル手段62を設け、制御手段20は、上位装置10からのアクセス要求時にデバイス管理テーブル手段62を参照してディスクアレイ手段18に対する処理を実行する。具体的には、デバイス管理テーブル手段62にミラーディスク手段36に設けたディスク装置の一方を現用とし、他方を予備用とする情報を登録する。

【0019】この場合、制御手段20は、リードアクセス時にデバイス管理テーブル手段62の参照でミラーディスク手段36の現用側のディスク装置からデータを読み出し、ライトアクセス時には現用および予備用の両方のディスク装置にデータを書込む。ミラーディスク手段36に対する切替は、ディスクアレイ手段18の切替制御手段40が行う。切替制御手段40は、制御手段20からのリード動作の指示に対しミラーディスク手段36の

現用側のディスク装置に切替えてデータを読み出し、ライト動作の指示に対しては現用および予備用の両方のディスク装置にデータを書込むように切替える。

【0020】また切替制御手段40を設ける代りに、制御手段20に対しミラーディスク手段36を構成する2台のディスク装置の各々を個別のポートを介して並列的に接続したディスクアレイ手段18とする。この場合、制御手段20はリード動作時には、ミラーディスク手段36の一方のポートによる現用側のディスク装置のアクセスでデータを読み出し、ライト動作時には、両方のポートによる現用および予備用の両方のディスク装置のアクセスでデータを書込むようにしてもよい。

【0021】この場合、制御手段20は、上位装置10からの書込要求時のデータストライピングとして、書込データをミラーディスク手段36の並列アクセス台数に基づいてストライピングした後に各ストライピングデータと同一データを生成する。このようにミラー化されたストライピングデータ（一対の同一データ）を、ミラーディスク手段36ごとの2つのポートに転送することで、同一データの書込みができる。

【0022】一方、ミラーディスク手段36を構成要素としたディスクアレイ手段18は、コストが増加し、また設置スペースも増加する。この点を解決するため、次のような実装構造をとる。まずディスクアレイ手段18は、着脱自在な少なくとも2台のディスクモジュール50を内蔵したディスクユニット46と、ディスクユニット46の収納部を複数備え、予め定めたアレイ構成に従って複数台の前記ディスクユニット46を実装した装置筐体44とで構成する。

【0023】ディスクユニット46は、着脱自在なディスクモジュール50に対する共通部として電源部および冷却手段を備える。またディスクモジュール50は、少なくともディスク媒体、ディスク回転機構、ヘッド機構を備える。さらにディスクユニット46に、制御手段20からのリード動作の指示に対しミラーディスク手段36の現用側のディスク装置に切替えてデータを読み出し、ライト動作の指示に対しては現用および予備用の両方のディスク装置にデータを書込むように切替える切替制御手段40の回路ユニットを内蔵させる。

【0024】この構成により、媒体サイズの異なる単一のディスクモジュールを内蔵した他のディスクユニットの収納を予定したディスクアレイ筐体に、本発明のディスクモジュールを実装してミラーディスク手段36を構成要素とするディスクアレイを簡単に実現できる。またアレイ構成要素のミラー化に伴い、ディスク装置の障害を可能な限り速かに知って対応することが信頼性を更に高めるために必要となる。

【0025】そこで本発明にあっては、ディスクアレイ手段12に対するアクセスの実行に対しチェック回路手段74において異常が発生したか否かを判定するアクセ

ス判定手段64と、アクセス判定手段64で判定した異常の回数を計数する異常回数計数手段66と、異常回数計数手段66の異常回数が所定の閾値以上となったときにデバイス障害を判定し、デバイス管理テーブル手段62に障害情報を登録する障害判定手段72を設ける。

【0026】またアクセス判定手段64の異常判定に対応して作成した予防保守情報を上位装置10で認識可能な状態に保持する予防保守情報格納手段68を設け、上位側でディスクアレイの状態を認識して保守が迅速にできるようにする。更に、制御手段20の空き状態（アイドル状態）で、ディスクアレイ手段18に対し擬似的なアクセス動作、すなわちシミュレーションを実行して各ディスク装置の状態を検査するデバイス検査手段70を設ける。

【0027】デバイス検査手段70は、例えばディスクアレイ手段18の特定データ領域（CE領域）に対しダミーデータを用いた擬似的な書込動作を実行する。このようなディスクアレイ手段18に対する擬似的なアクセス動作でアクセス判定手段64が異常を検出すると、異常検出ごとに異常回数係数手段66で異常回数を計数させ、異常回数が所定の閾値を越えたことを障害判定手段72で判定すると、障害デバイスとしてデバイス管理テーブル手段62に障害情報を登録する。

【0028】さらに、デバイス検査手段70による擬似的なアクセスでアクセス判定手段64が判定した異常情報に基づき予防保守情報を生成し、この予防保守情報を上位装置10で認識可能な状態に保持する予防保守情報格納手段68を設ける。デバイス管理テーブル62に障害情報が登録された状態では、制御手段20は、リード動作の実行時に前記デバイス管理テーブル手段62の障害情報を参照して障害ディスク装置を認識した場合、障害ディスク装置が現用側であれば予備用からのリード動作に切替え、障害ディスク装置が予備用であれば、現用からのリード動作を維持する。

【0029】また制御手段20は、ライト動作の実行時に前記デバイス管理テーブル手段62の障害情報を参照して障害ディスク装置を認識した場合は、障害ディスク装置をライト動作の対象から切り離す。この障害ディスク装置の切替は、デバイス管理テーブル62に障害情報の登録に伴ない、現用と予備用を示す識別情報を切替えることによって実現できる。

【0030】更に本発明は、アレイ要素をミラーディスク手段で構成すると同時に冗長情報も格納し、ミラード化と冗長化を複合した信頼性をさらに向上する。すなわち、同一データを格納する2台のディスク装置36を備えたミラーディスク手段36を並列アクセス可能に複数配置したディスクアレイ手段18と、上位装置10からの書込データをストライピングしてパリティデータと共にディスクアレイ手段36に並列に書込み、ディスクアレイ手段36からの並列読出データを合成して前記上位

装置10に転送する制御手段20とで構成する。

【0031】さらに本発明は、通常は単なる並列ディスク装置を備えたディスクアレイとして機能するが、その中の1台に障害発生の可能性ありと予測的に判定された場合に、予備のディスク装置とのミラード化を動的に行い、実際のディスク故障時に予備側への切替が瞬時にできるようにする。このため、複数のディスク装置32を並列アクセス可能に配置したディスクアレイ手段18と、ディスクアレイ手段18の同一ランクに属する少なくとも1台のディスク装置を予備用に割当てと共に他の複数ディスク装置をデータ格納用に割当て、上位装置10からの書込データをストライピングしデータ格納用ディスク装置に並列的に書込み、前記データ格納用ディスク装置からの並列読出データを合成して上位装置10に転送する制御手段20を設ける。

【0032】更に、アクセス判定手段64で判定したデータ格納用ディスク装置の異常回数を異常回数係数手段66で計数し、この計数結果が通常の障害半手用いる閾値より低い所定の閾値を越えたときに、近い将来、障害がは発生することを予測的に判定する障害判定手段72を設け、障害判定手段72で障害を予測判定した時に、構成制御手段により予備用ディスク装置を選択してミラード化する。

【0033】すなわち異常回数が障害と判定される値に達する前の状態でブリアラーム的に障害を予測判定し、この段階で構成制御手段は、障害判定ディスク装置の格納データを複写し、複写後に障害判定ディスク装置と予備用ディスク装置でミラーディスク手段36を構成し、制御手段20に両方のディスク装置への同時データの書込といずれか一方のディスク装置からのデータ読出をセットする。

【0034】ここで、構成制御手段は、予備用ディスク装置を現用にセットして制御手段20によるデータ読出しを行わせる。また障害判定手段72は、障害判定ディスク装置の修理交換を促すため、障害の予測的な判定が行われたことを外部に出力表示させる手段を備える。

【0035】

【作用】このような本発明のディスクアレイ装置によれば、ディスクアレイ装置の各構成要素を2台のディスク装置を1組として同一データを格納するミラーディスク構成とすることで、冗長情報を必要とすることなく、いずれかのディスク装置の故障によるデータ喪失に対し瞬時に対応することができ、きわめて高い信頼性を確保することができる。

【0036】また1つのディスクユニットに2台のディスクモジュールを着脱自在に実装可能としたことで、通常は1台のディスクモジュールの実装によるアレイ構成であっても、残り1台を追加実装することで、ミラーディスク構成に簡単に変更できる。更に、3.5インチ・ディスクモジュールを2台実装可能なディスクユニット



を、既存の5インチ・ディスクモジュールを1台備えたディスクユニットと同一ユニットとしておくことで、5インチ・ディスクユニットを用いたディスクアレイを、3、5インチ・ディスクモジュールを2台備えたディスクユニットに交換することで、簡単にミラード化された本発明のディスクアレイ装置を構築できる。

【0037】一方、ミラーディスクの一方を現用、他方を予備用に割当て、通常は、データ書込は現用と予備用の両方に対し行い、データ読出しは現用から行っているが、統計的に求めたシークエラー、リードエラー、ライトエラーなどの異常発生の数値を監視し、現用のディスク装置のエラー回数が所定の閾値を越えた場合には、故障の可能性が高いと判定し、故障が発生する以前に、現用と予備用との割当てを切替える。

【0038】このため実際に現用がディスク故障となった場合にも、既に予備用への割当てが済んでいるため、予備用ディスク装置の故障となり、特にリードアクセスに対する処理を全く中断することなく、障害に対処できる。また障害判定結果を予防保守情報として上位装置で常に参照可能に保持しているため、上位装置側でディスクアレイの障害状況が出力表示でき、オペレータ又は保守要員が迅速且つ適切に障害判定ディスク装置に対する保守作業を行うことができる。

【0039】またディスクアレイの状態を上位装置からのアクセスがない空き状態を利用して制御装置側が書込シミュレーションによる検査処理を実行し、積極的にディスク装置の状態を検査し、速かに故障の可能性のあるディスク装置を認識して対応することができる。このようなミラード化されたディスクアレイ装置の信頼性をさらに向上するためには、ミラード化されたディスクアレイに対し冗長情報としてパリティをミラード化して格納するRAID3あるいはRAID5の動作形態を取ればよい。

【0040】更に、通常は、単なる並列アクセスであり、障害と判定される少し前にミラーディスク構成に切替えるダイナミックなミラード構成とする。このため同一ランクに少なくとも予備用のディスク装置を1台割当てておき、通常は、予備以外のディスク装置の並列アクセスを行う。運用中にいずれかのディスク装置のエラー回数が、予め定めた通常の障害判定より低めの閾値を越え、近い将来、障害と判定される状況に至ったら、予備用ディスク装置を選択し、障害判定対象となったディスク装置のデータを複写する。

【0041】このようにディスク装置間でのデータ複写は、通常の障害判定より早い段階で行われるため、ディスク装置の障害で複写不能となってしまう事態を極力回避できる。このプリアラーム的な障害判定は、通常の障害判定の閾値の半分というように、かなり早い段階で行うことが望ましい。この予備ディスク装置へのデータ複写は、アイドル状態で行えばよい。予備へのデータ複写

が済んだならば、障害判定ディスク装置と予備用ディスク装置でミラーディスクを構成し、以後、同一データの書込みを行う。またデータの読出しは、予備用ディスク装置を現用に割当てて行う。

【0042】このような障害予測判定時の動的なミラード化により、必要最小限のディスク装置で、実質的なミラーディスク構成のディスクアレイを実現できる。

【0043】

【実施例】図2は本発明のディスクアレイ装置が適用される計算機の入出力サブシステムのハードウェア構成を示す。図2において、上位装置としてのホストコンピュータ10にはチャンネル装置14が設けられ、チャンネル装置14に対しチャンネルインタフェースバス16を介してコントローラ12を接続している。

【0044】チャンネルインタフェースバス16としては、BMCインタフェース（ブロック・マルチプレクサ・チャンネルインタフェース）やSCSIを使用することができる。コントローラ12には入出力デバイスとしてディスクアレイ18が接続されている。

【0045】ディスクアレイ18は並列アクセス数を示す横方向に並んだポートP0～P3と、横方向のP0～P3の並びを縦方向に多段構成したランクR0～R3の構成をもち、ポートP0～P3とランクR0～R3で定まる位置にディスクアレイの構成要素として、本発明にあっては、ミラーディスク装置36-1～36-34を配置している。

【0046】ミラーディスク装置36-1～36-34は例えばポート番号P0、ランク番号R0のミラーディスク装置36-1を例にとると、2台のディスク装置32-1、32-2を接続しており、それぞれをコントローラ12側に設けたミラード制御部30-1からのデバイスバスに接続している。このようなミラーディスク装置36-1の構成は他のミラーディスク装置36-2～36-34についても同じである。

【0047】ミラーディスク装置36-1はコントローラ12により同一データの書込みを受ける。従って、ディスク装置32-1、32-2には常に同一データが格納されている。一方、データ読出しはいずれか一方のディスク装置32-1、32-2から行えばよい。本発明にあっては、コントローラ12において予めミラーディスク装置36-1の左側のディスク装置32-1を現用に割り当て、右側のディスク装置32-2を予備用に割り当てている。

【0048】このため、データ書込みは現用および予備用のディスク装置32-1、32-2に対し行われるが、データ読出しは現用のディスク装置32-1のみから行われることになる。コントローラ12にはMPU20が設けられ、MPU20の内部バス34にチャンネルインタフェース制御部22、制御記憶部24、データバッファ26およびデバイスインタフェース制御部28を接



続している。

【0049】MPU20は後の説明で明らかにするように、プログラム制御に従ってディスクアレイ18のミラーディスク装置群を対象としたホストコンピュータ10からの要求に基づく処理動作を実行する。尚、この実施例において、ディスクアレイ18側は4つのポートP0～P3についてディスク装置を実装した場合を示しており、ミラー制御部30-5についてディスク装置は空き状態となっている。

【0050】図3は図2のコントローラ12に設けたミラー制御部30-1の詳細をディスクアレイ側と共に示した実施例構成図である。図3において、コントローラ12側に設けたミラー制御部30-1には切替制御部38、ディスク制御部40-1、40-2が設けられている。ディスク制御部40-1、40-2のそれぞれからはデバイスインタフェースバス42-1、42-2が引き出され、この実施例にあっては2台でミラーディスク装置36-1～36-31を構成する磁気ディスク装置をランクごとに2台ずつ接続している。

【0051】切替制御部38はMPU20側で予め設定したミラーディスク装置の動作モードの割当て、即ち現用割当てと予備用割当てに従った切替処理を行う。ここで例えばディスク制御部40-1側を現用、ディスク制御部40-2側を予備用に割り当てていたとすると、MPU20からのライト動作の指示に対してはライト動作命令をディスク制御部40-1、40-2の両方に供給する。

【0052】同時に指示されたデバイスIDで特定される例えばランクR0のミラーディスク装置36-1におけるディスク装置32-1、32-2に対する書込動作を行わせる。一方、MPU20側から読出動作を指示された場合には現用となるディスク制御部40-1側のみに切り替え、同時に、指定されたデバイスIDに対応する例えばミラーディスク装置36-1の現用側のディスク装置32-1からの読出動作を行う。

【0053】このような図3に示すミラー制御部30-1の構成は図2に示した残りのミラー制御部30-2～30-5についても同様である。図4は本発明に用いるディスクアレイの筐体構造の実施例を示した実施例構成図である。図4において、ディスクアレイ筐体44は、この実施例にあっては5段2列の収納部を形成しており、各収納部に図示のように10台のディスクユニット46-1～46-10を実装している。

【0054】ディスクアレイ筐体44に設けられた開閉自在な扉45には内蔵した2列5段構成のディスクユニット46-1～46-10に対応して5箇所に通風口が設けられており、通風口の内側に図示のようにフィルタ48-1～48-5を装着している。図4のディスクアレイ筐体44は10台のディスクユニット46-1～46-10を全て実装した最大構成状態を示しているが、

必要に応じた台数のディスクユニットを実装することができる。

【0055】図5は図4のディスクアレイ筐体44に実装したディスクユニット46-1を取り出して示す。図5において、ディスクユニット46-1内には、この実施例にあっては2台の3.5インチ・ディスクモジュール50-1、50-2が実装されている。即ち、ディスクユニット46-1のケース内の中央にはモジュール実装基板54が装着されており、ケース背後が開口している。

【0056】この背後の開口部より図示のようにモジュール回路基板52-1、52-2のそれぞれに組み付けられたディスクモジュール50-1、50-2を差し込んでコネクタにより接続することで、動作可能状態に組み込むことができる。またモジュール実装基板54の右側には電源ユニット56と冷却ファン装置58が組み込まれている。

【0057】ここでディスクユニット46-1は3.5インチ・ディスクモジュール50-1、50-2の2台を実装しているが、このユニット形状は例えば1つ媒体サイズが上の5インチ・ディスクモジュールを1台内蔵したディスクユニットと同一形態および構造としている。また容量的にも3.5インチ・ディスクモジュール50-1、50-2のそれぞれは同じユニットサイズの5インチ・ディスクユニットと同じ容量を有する。

【0058】従って、3.5インチ・ディスクモジュール50-1、50-2のいずれか一方の1台のみの実装状態で従来の5インチ・ディスクユニットを用いたディスクアレイ筐体の実装すれば、そのまま3.5インチ・ディスクユニットを用いたディスクアレイを実現できる。一方、図5に示すように、2台の3.5インチ・ディスクモジュール50-1、50-2を実装した状態で従来の5インチ・ディスクユニットに置き換えれば、従来の5インチ・ディスクユニットを3.5インチ・ディスクモジュール2台を実装したディスクミラー装置に簡単に変更することができる。

【0059】更に、3.5インチ・ディスクモジュールの実装数とユニット数を適宜に決めることで、入出力サブシステムの性能に見合ったディスクミラー装置を構成要素とする場合を含む適宜のディスクアレイ構成を実現することができる。図6は本発明の処理機能を示した実施例構成図である。図6において、コントローラ12に設けられたMPU20のプログラム制御およびファームウェアによって、ディスクアレイ制御部60、デバイス管理テーブル62、アクセス制御部64、異常回数計数部66、予防保守情報格納部68、障害判定部72、データ転送制御部75およびチェック回路部74としての機能が実現される。

【0060】コントローラ12に接続したディスクアレイ18としては、説明を簡単にするため、1ランク分の

ミラーディスク36-1~36-4のみを示している。まずディスクアレイ18はデバイス管理テーブル62に基づいて各種の動作を行い、デバイス管理テーブル62は例えば図7に示す構成をもつ。図7において、デバイス管理テーブル62はミラードID、デバイスID、ランク番号およびポート番号の4つのパラメータによってディスク装置を特定することができる。

【0061】ミラードIDは例えばランクR0を例にとると、ミラーディスク装置36-1~36-4に予め割り当てられた識別番号であり、この例では00、01、02、03で表わしている。またデバイスIDはディスクアレイ18におけるデバイス物理番号であり、例えばランクR0の8台のディスク装置を例にとると、物理デバイスID「00」~「07」で表わしている。更にランク番号はディスクアレイ18のランク位置R0~R3を示すもので、例えばランク番号R0については「00」で示している。

【0062】次のポート番号もディスクアレイ18におけるミラーディスク単位のポート番号であり、例えばR0を例にとると、ポート番号「00」~「03」で表わしている。このポート番号に続いて、ディスクアレイ18に設けたディスク装置の状態を示す現用フラグ、予備用フラグ、異常回数および故障フラグの各情報が登録されている。

【0063】現用フラグおよび予備用フラグは電源投入によるシステム立上げ時に初期化され、図7のようにミラードIDで指定されるミラーディスク装置の先のデバイスIDに現用フラグ1がセットされ、次のデバイスIDに予備用フラグ1がセットされている。また異常回数は立上げ時には0にリセットされ、同様に故障フラグも立上げ時には0にリセットされている。

【0064】異常回数は図6に示した異常回数計数部66の計数結果を格納している。すなわちコントローラ12に設けたアクセス判定部64は、ディスクアレイ制御部60でディスクアレイ部18に対しリード動作又はライト動作を実行した際のチェック回路部74によるチェックにおいてシークエラー、リードエラー、ライトエラーなどエラーが発生したか否かを判定している。アクセス範囲部64でエラーが判定されると、異常回数計数部66の計数値が1つカウントアップされ、カウントアップした値をデバイス管理テーブル62上の異常回数として更新する。

【0065】また障害判定部72が異常回数を異常回数計数部66で計数するごとに、予め定めた障害判定の閾値と比較しており、異常回数が障害判定の閾値に達すると、障害発生と判定し、デバイス管理テーブル62の故障フラグを1にセットする。図8はミラードID=00をもつミラーディスク装置36-1の現用のディスク装置32-1の異常回数が10回となり、閾値に達して障害ディスクと判定することで故障フラグが1にセットさ

れた状態を示している。

【0066】更に図6のコントローラ12には予防保守情報格納部68が設けられている。この予防保守情報格納部68にはアクセス判定部64でエラー判定が行われるごと、エラー発生デバイスを示すデバイスIDとエラー種別が予防保守情報として格納される。予防保守情報格納部68に格納された予防保守情報は、障害範囲部72による障害判定の通知以前に、ホストコンピュータ10側から常に参照することができる。

【0067】例えばホストコンピュータ10からコントローラ12に対し任意の入出力要求が行われたときのステータス情報の中に予防保守情報を含めて送り返すことで、ホストコンピュータ10側で予防保守情報の存在を認識し、ディスプレイやプリンタに取得した予防保守情報を出力表示する。このため、オペレータや保守要員はコントローラ12側でディスク装置のエラー発生の状況をリアルタイムで把握できる。

【0068】更にコントローラ12側にはデバイス検査部70が設けられている。デバイス検査部70はホストコンピュータ10からのアクセスのない空き状態、即ちコントローラ12のアイドル状態で起動し、ディスクアレイ18に対しダミーデータを用いた書込動作のシミュレーションを実行する。このため、ディスクアレイ18の各ディスク装置には予めデバイス検査部70によるダミーデータの書込みのシミュレーションに使用される、通常CE領域として知られた固有の検査領域が確保されている。

【0069】デバイス検査部70によるデータ書込みのシミュレーションについても、アクセス判定部64がエラー判定を行っており、エラーが検出されると異常回数計数部66の計数値を1つカウントアップする。また異常回数計数部66の計数値が所定の閾値に達すると障害判定部72が障害を判定し、ホストコンピュータ10の要求処理時と同様、デバイス管理テーブル62に対する故障フラグのセットを行う。勿論、アクセス判定部64でデバイス検査部70により擬似的なアクセス動作でエラーが判定されると、保守情報格納部68に対しホストコンピュータ10から参照可能な予防保守情報の格納がその都度行われる。

【0070】更にデータ転送制御部72はディスク制御部60による制御のもとに、ホストコンピュータ10からのアクセス要求に従ったディスクアレイ18とホストコンピュータ10との間のデータ転送を行う。次にディスクアレイ制御部60によるミラーディスク装置を構成要素としたディスクアレイ18に対する処理動作を説明する。

【0071】図9はコントローラ12のディスクアレイ制御部60で行われる書込データのストライピングと、ディスクアレイ18側のデータ格納状態を示している。図9において、ホストコンピュータ10がアクセスする

データ単位を論理ブロック（ホストブロック）76で示しており、例えばディスク装置の1セクタを512バイトとすると、論理ブロック76は2セクタ即ち1,024バイトのサイズをもつ。

【0072】このようなホストコンピュータ10からの書込データとしての論理ブロック74を受けると、ディスクアレイ18の並列アクセス数の整数倍で割った数のストライプデータ78に分割する。この例では並列デバイス数4の4倍となる16分割した場合を示している。このため、1,024バイトの論理ブロック74を16

分割すると1つのストライプデータのサイズは64バイトとなる。

【0073】即ち、論理ブロック76を16分割したストライプデータ76をストライプデータL00～L15で示している。論理ブロック74のストライピングが済むと、ストライプデータ78を並列デバイス数単位に取り出してミラーディスク装置36-1～36-4に対し並列的に書き込む。この並列書き込みを4回繰り返すことで、1論理ブロック分のデータをミラーディスク装置36-1～36-4に格納することができる。

【0074】これをミラーディスク装置36-1～36-4の中の1台のディスク装置について見ると、64バイトのデータが4つ置きに分散して2セクタ分格納された状態となる。このようなデータストライピングおよびデータの格納はRAID3の動作モードと基本的に同じであり、パリティデータを計算して格納していない点だけが相違する。

【0075】またミラーディスク装置36-1～36-4にあっては、それぞれ現用と予備用の2台のディスク装置で構成されており、それぞれに同じデータを格納していることが明らかである。図10は図6に示したコントローラ12による本発明のディスクアレイ装置の処理動作を示したフローチャートである。

【0076】図10において、まず装置の電源を投入してシステムを立ち上げると、ステップS1でデバイス管理テーブル62の初期化が例えば図7に示したように行われる。続いてステップS2でホストコンピュータ10からのアクセス要求の有無をチェックしており、アクセス要求のないアイドル状態ではステップS15のデバイス検査処理を行っている。アクセス要求があるとステップS3で指定されたデバイスIDとなるミラーディスク装置に対しセットアップ処理を行う。

【0077】このコントローラ12からのセットアップ要求に対し、指定されたミラーディスク装置の各ディスク装置からは正常応答または異常応答を返す。ミラーディスク装置を構成する2台のディスク装置が共に正常であればステップS4でデバイスエラーなしとしてステップS5に進み、アクセス要求がリード要求かライト要求かを判別する。

【0078】リード要求についてはステップS6に進

み、ミラーディスク装置の現用側からのリード処理を実行する。一方、ライト要求であればステップS7に進み、現用および予備用のディスク装置に同一データを書き込むライト処理を実行する。ステップS8にあっては、ステップS6のリード処理またはステップS7のライト処理が正常終了したか否かチェックしており、正常終了であればステップS9でリード処理またはライト処理においてリトライで回復したシークエラー、ライトエラー、リードエラーのいずれかがあるかを判定する。

【0079】いずれのエラーもなければ再びステップS2に戻って、次のホストコンピュータ10からのアクセス要求を待つ。ステップS9でリード処理中またはライト処理中にリトライで回復したシークエラー、リードエラーまたはライトエラーがあった場合にはステップS10に進み、異常回数計数部66でエラー回数を1つカウントアップする。

【0080】続いて障害判定部64が現在のエラー回数を所定の閾値と比較する。閾値に達していなければステップS16で、このときのエラーに関する予防保守情報を予防保守情報格納部66にセットしてステップS2に戻る。もしエラー回数が閾値に達していた場合にはステップS12に進んで、近い将来、ハードウェアエラーなどの故障により動作不能となことから、デバイス障害を判定してデバイス管理テーブル62上に故障フラグをセットする。

【0081】例えば、図8のデバイスID番号00のディスク装置32-1に示すように、異常回数が10回となって閾値に達すると故障フラグ1のセットが行われる。このように故障フラグのセットが行われると、次のステップS13でもし故障フラグのセットが現用のディスク装置に対し行われていた場合には、ステップS1の初期化処理で割り当てた現用フラグの予備用フラグへの代替を行う。

【0082】例えば図7に示すように、デバイスID番号00のディスク装置32-1は現用フラグが1にセットされて最初、現用となっていたものが、図8に示す故障フラグ1のセットで現用フラグを0にリセットし、それまで予備用として予備用フラグが割り当てられていたデバイスID番号01のディスク装置32-1の予備用フラグを0にリセットして現用フラグを1にセットする。

【0083】この実施例では、故障フラグ1をセットしたそれまでの現用ディスク装置36-1について、現用フラグを0にリセットすると同時に予備用フラグも0にリセットしているが、並行してライト動作を継続したい場合には予備用フラグを1にセットすればよい。このようなデバイス管理テーブル62における現用フラグと予備用フラグの入替えに対し、その後の処理においてディスクアレイ制御装置60は現用フラグが1にセットされた側を現用、予備用フラグが1にセットされた側を予備

用としてライト処理またはリード処理を実行するようになる。

【0084】従って、障害フラグがセットされたそれまでの現用のディスク装置はリード処理の対象から除外され、新たに現用となった予備用のディスク装置からのリード処理に切り替わる。またライト処理については現用および予備用の如何に関わらず、同時にデータ書込みが行われる。しかし、図8に示すように予備用フラグも0にしていた場合には、それまでの予備用で現用に切り替わったディスク装置に対してのみデータ書込みが行われ

る。

【0085】再び図10を参照するに、ステップS13で現用から予備用への代替が済むと、ステップS14でホストコンピュータ10に障害は発生を通知し、障害と判定されたディスク装置の修理交換をオペレータや保守要員に促す。一方、ステップS3のセットアップ処理に対し、デバイスIDで指定されたディスク装置がハードウェア故障を起こし、全く動作できずにデバイスエラーを判定した場合には、直ちにステップS12に進み、ステップS16の予防保守情報のセットで、ハードウェアエラーがホストコンピュータ10側から参照できるようにする。この場合、もし現用のディスク装置の障害であれば、予備用への割当ての代替を行い、更にステップS16で予防保守情報を格納してホストコンピュータ10側でデバイスエラーを参照可能とする。

【0086】またデバイスエラーの場合は、ステップS11のエラー回数の判定に基づく障害判定とは異なり、回復不可能な故障であることから、予防保守情報にディスク装置の機能停止である旨の識別表示をつけ、ソフトウェアによる予防保守情報と区別できるようにする。図11のフローチャートは図6のステップS15に示したコントローラ12のデバイス検査部70によるデバイス検査処理を示す。

【0087】図11において、まずステップS1でコントローラ12がアイドル状態か否かチェックしており、アイドル状態にあるとステップS2以降の処理を行う。ステップS2にあつては、所定の順番に従った指定ランクの現用および予備用のディスク装置に設けられた保守用のデータ領域に対し、ダミーデータの書込みを検査シミュレーションとして実行する。

【0088】このダミーデータの書込みにおいて、シークエラー、ライトエラーなどのエラーが発生すると、ステップS3でエラーありを判定し、ステップS4で異常回数計数部66を使用して、それまでのエラー回数を1つカウントアップする。続いてステップS5でエラー回数を所定の閾値と比較し、閾値に達していなければ、ステップS10で予防保守情報をセットしてステップS1に戻る。エラー回数が閾値に達していればステップS6で故障フラグをデバイス管理テーブル62にセットする。

【0089】続いてデバイス管理テーブル62において、現用のディスク装置については予備用のディスク装置への代替を行うためのフラグ割当ての入替えを行う。続いてステップS8でホストコンピュータ10に対し障害発生通知を行う。以上の処理を終了するとステップS9で次のランクを指定し、再びステップS1のアイドル状態のチェックに戻る。

【0090】このような検査処理により例えばオンラインシステムにあつては、処理負荷の少ない夜間などの時間帯においても、自動的にミラーディスク装置を構成するディスクアレイのディスク装置に対する書込シミュレーションが一定の検査周期ごとに実行され、常にディスク装置の状態を監視し、可能な限り早めに、故障する可能性のあるディスク装置を予測判定することができる。

【0091】図12はコントローラ側における障害予測判定あるいはデバイス故障判定に基づく予防保守情報や障害情報の出力表示で、保守要員が障害対象となったディスク装置をディスクアレイ筐体から外して修理し、正常なディスク装置を筐体の実装した後に行う復旧処理のフローチャートを示す。図12において、まずステップS1で正常なディスク装置の筐体への組み込み完了に伴う保守要員のホストコンピュータからのコマンドあるいは保守パネルからのスイッチ操作による復旧指示の有無をチェックしている。

【0092】復旧指示を判別するとステップS2に進み、コントローラ12がアイドル状態か否かチェックする。アイドル状態にあるとステップS3に進み、復旧によりミラーディスク装置の一方のディスク装置として組み込まれた復旧側ディスク装置に対し、現用のディスク装置のデータを複写する処理を開始する。

【0093】次のステップS4にあつては、予備固定モードか否かチェックしている。即ち、ミラーディスク装置における2台のディスク装置に対する現用と予備の割当てを固定的に定めていた場合には、データの複写が完了した時点でデバイス管理テーブル62の現用フラグおよび予備用フラグをシステム立ち上り時の初期化時と同じ状態に初期化する。

【0094】一方、現用と予備用を固定せずに動的としていた場合には、ステップS5におけるデバイス管理テーブル62の初期化は行わず、現在、機能しているディスク装置を現用のまま維持し、復旧により組み込んだディスク装置を予備用とする。更にステップS6において、予防保守情報のクリア、およびデバイス管理テーブルの故障フラグおよび異常回数のクリアを行う。

【0095】図13はミラーディスク構成のディスクアレイを使用してデータの並列アクセスと同時に冗長情報としてパリティを同時に格納するようにしたことを特徴とする。即ち、ディスクアレイのミラード化により信頼性は極めて向上するが、更にミラード化に冗長情報の格納を加えることで極めて高い信頼性を実現することがで

きる。

【0096】図13にあっては、データ用のミラーディスクの構成は図2の実施例と同じであるが、パリティ格納用にミラード制御部30-5よりのポートP4にミラーディスク装置36-5～36-35を4ユニット設けている。図14は図13のパリティ格納を行うミラーディスク構成のディスクアレイにおけるRAID3の動作モードにおける処理動作を示す。

【0097】図14のRAID3における論理ブロック76のストライピングは、図9に示したパリティを格納しない場合と同じであるが、新たに設けたパリティ用のミラーディスク装置36-5に対し並列デバイス数4単位の4つのストライピングデータ78、例えばストライピングデータL00～L03からパリティP00を計算して、ミラーディスク装置36-5の現用および予備用のディスク装置に同時に格納するようにしている。

【0098】このようなパリティを格納するRAID3の動作形態とした場合には、ミラーディスク装置36-1～36-5のいずれかを構成する現用と予備の2台のディスク装置が同時に故障してデータが失われた場合にも、残りの正常なミラーディスク装置のデータから失われたデータを復元することができる。図15は図13についてRAID5の動作形態によるデータのストライピングとデータの格納状態を示す。

【0099】図15において、まず論理ブロック76はディスク装置の例えば1セクタ(512バイト)のサイズであり、この場合には論理ブロックL0～L5の6つが書込データとして提供された状態を示している。このような書込データにつき、論理デバイス数4に対応した4つの論理ブロック例えば論理ブロックL0～L3をストライピングし、各論理ブロックL0～L3のそれぞれをミラーディスク36-1～36-4に並列的に格納する。

【0100】同時に論理ブロックL0～L3からパリティP0を計算し、このセクタについてはミラーディスク装置36-5に格納する。RAID5にあっては、セクタごとにパリティの格納位置がP0、P1、P2、P3、P4に示すように順次変化するようになる。このRAID5のミラーディスクを構成要素としたディスクアレイの動作形態にあっても、1つのミラーディスクを構成する現用と予備の2台のディスク装置が同時に故障しても、他の正常なミラーディスク装置からデータを修復することができる。

【0101】図16は図13に示したパリティ格納を行うミラーディスク装置によるディスクアレイの構成について、各ミラーディスク装置の予備側のディスク装置を取り外して通常のRAID構成のディスクアレイ装置とした場合を示している。ここで本発明にあっては、図5に示した2台の3.5インチ・ディスクモジュール50-1、50-2を実装したディスクユニット46-1の

ディスクアレイ筐体に対する組込みでミラーディスク装置を構成要素としたディスクアレイを構成している。

【0102】このため、ディスクユニットに実装している1台のディスクモジュールを外すことで簡単に通常のRAID構成の実装状態を実現できる。逆に図16の通常のRAID構成について、ディスクモジュールを追加することで図2あるいは図13に示したミラーディスク構成のディスクアレイを簡単に実現できる。

【0103】図17はミラーディスク装置を構成要素とする本発明のディスクアレイ装置の他の実施例を示した実施例構成図であり、この実施例にあってはコントローラ12からの独立したポートに接続しているディスク装置を組み合わせてミラーディスク装置として動作するようにしたことを特徴とする。即ち、図2の実施例にあっては、コントローラ12はデバイスインタフェース制御部28に対しミラード制御部30-1～30-5を設け、ミラード制御部については図3に示すように、切替制御部38によって切り換えることでミラーディスク装置としての機能を実現している。

【0104】これに対し図17の実施例にあっては、コントローラ12のデバイスインタフェース制御部28に対し個別にディスク制御部40-1～40-8を設けて独立にディスクアレイ18のポートP0～P7を構成し、各ポートに4ランク分のディスク装置32-1～32-38を接続している。このようなディスクアレイ18の独立したポートP0～P7に対し、図7に示したようなデバイス管理テーブル62におけるミラードID番号、デバイスID番号の割当てを行うことで、結果としてミラーディスク装置36-1～36-34をディスクアレイ18に割り付けることができる。

【0105】図18は図17の独立したポートをもつディスクアレイ18のミラーディスク装置に対するデータストライピングとデータ格納状態を示している。図18において、コントローラ12はディスクアレイ18がミラーディスク構成か否かは特に意識しておらず、論理ブロック76からストライピングしたデータをデバイスIDで指定されるディスク装置に転送するに過ぎない。

【0106】論理ブロック76からのデータストライピングについて、この場合には論理ブロック76を例えば16分割してストライピングデータL00～L15を得る。同時に、同じデータL00～L15を生成し、図示のようにL00～L15を2つずつ並べたダブルストライプデータ80を生成する。このダブルストライプデータ80の生成に対し、ミラーディスク装置を意識せずに並列デバイス数8台分のダブルストライプデータ例えばL00、L00、・・・L03、L03を取り出して並列的に書き込むことで、結果的にミラーディスク装置と同じデータ格納状態を実現することができる。

【0107】このように、コントローラ12側でミラーディスク装置固有の切替制御機構をもたなくとも、通常

のRAID構成のコントローラのポート数を増加させるだけで簡単にミラーディスク装置を構成要素としたディスクアレイ装置が実現できる。図19は本発明のディスクアレイ装置の他の実施例を示したもので、ディスク装置の障害予測判定時に動的にミラーディスク装置を構成するようにしたことを特徴とする。

【0108】ディスクアレイの構成要素をミラーディスク装置としたディスクアレイ装置の最大の欠点は、ディスクアレイに使用するディスク装置が通常の2倍に増加し、コストアップになってしまう点である。しかしながら、磁気ディスク装置の小型化と低価格化に伴い、この問題は必ずしも妨げとはなっていないが、ディスクアレイに使用するディスク装置の台数を少なくすることは依然として望まれる。

【0109】図19の実施例にあっては、コントローラ12に設けたディスク制御部40-1~40-5に対応してディスクアレイ18にポートP0~P4を設け、4段のランク構成を例にとっている。ディスクアレイ18のポートP0~P3はデータ格納用であり、各ランクごとに単一のディスク装置を接続している。

【0110】これに対し、ポートP4には予備用のディスク装置32-5、32-10、32-15、32-20を接続している。通常はディスクアレイ18のランクごとにポートP0~P3を対象に並列的なデータ読み書きを行っている。この状態でいずれかのディスク装置のエラー回数が閾値に達して障害が判定されると、同一ランクに存在する予備用ディスク装置の使用によるミラーディスク装置の構成制御が実行される。ここで予備ディスク装置の使用によるミラーディスク装置の構成制御には、現用から予備へのデータ複写が必要であり、現用について通常の障害判定が行われた場合には、データ複写ができないことから、障害判定に使用するエラー回数の閾値を低めに設定し、早い段階で障害を予測的に判定してミラーディスク装置への構成制御を行う。この構成制御を開始するための閾値としては、例えば通常のエラー回数の障害判定に用いる閾値の半分程度の値でよい。例えば通常、10回で障害を判定していたならば、この場合は5回の閾値で早めに障害を判定する。

【0111】このような構成制御によつて、予測的に障害と判定されたディスク装置は予備用ディスク装置との組合せでミラーディスク装置として、以後、動作することとなり、その後に通常の閾値による障害判定となった場合、直ちにミラーディスク装置の一方である予備用のディスク装置による正常なデータアクセスに切り替えることが瞬時にできる。

【0112】図20は図19における予測障害判定時のミラーディスクへの構成制御を示したフローチャートである。図6に示した障害判定部64あるいはデバイス検査部70により適宜のディスク装置で障害予測判定が行われると、図20の処理が開始される。まずステップS

1で障害対象として判定されたディスク装置と同一ランクの予備用ディスク装置を選択し、ステップS2でコントローラ12がアイドル状態か否かチェックする。

【0113】アイドル状態であればステップS3で障害判定ディスク装置のデータを、選択された予備用ディスク装置に複写する処理を開始する。複写中にホストコンピュータ10からリード要求又はライト要求があった場合は、複写処理を一旦中止し、リード又はライト動作を優先的に実行する。またその後の複写処理の再開は、一定時間を越えてアクセス要求がなかったときに再開する。これはディスクキャッシュにおけるキャッシュ書込後の空き時間を使用してディスク装置への書込みを行うライトバック処理に類似する。

【0114】ステップS4で予備用ディスク装置に対するデータの複写が終了すると、ステップS5で障害判定されたディスク装置とデータ複写が済んだ予備用ディスク装置を1つのミラーディスク装置とするデバイス管理テーブル62に対するセッティングが行われ、通常処理に復旧する。従って、この動的なミラーディスク構成制御が行われた以降については、予測障害が判定されたディスク装置と予備用ディスク装置で構成されるミラーディスク装置をディスクアレイの構成要素の1つに含んだ処理動作が行われる。

【0115】このため、万が一、障害予測に続いて実際にハードウェアエラーで故障が起きても、ミラード化された予備用ディスク装置による切替えで直ちに上位装置からの要求に対応することができる。また、予測障害の判定結果は予防保守情報としてホストコンピュータ10側で見ることができると、予測障害の判定対象となったディスク装置の修理交換を速やかに行うことができる。

【0116】この修理交換においても、予備用ディスク装置によるミラーディスク装置の構成制御が完了していると、特別なディスク装置の代替などを行うことなく予測障害の対象となったディスク装置を取り外して修理交換することができる。尚、図19の実施例にあっては、同一ランクに予備のディスク装置を1台設けた場合を例にとっているが、予備のディスク装置を2台とすれば同一ランクの2台のディスク装置の同時故障に対し瞬時に要求データの対応状態に切り替えることができる。

【0117】このように図19の実施例にあっては、通常のRAIDの形態で動作するディスクアレイに設けている予備ディスク装置の場合と同等なディスク台数で、実質的にディスクアレイの構成要素をミラーディスク装置としたと同じ極めて高い信頼性のディスクアレイ装置を実現することができる。また上記の実施例は4ポート、4ランクのミラーディスク装置を構成要素としたディスクアレイを例にとるものであったが、ポート数およびランク数は必要に応じて適宜に定めることができる。

【0118】更に本発明は実施例に示した数値による限



定は受けない。

#### 【0119】

【発明の効果】以上説明してきたように本発明によれば、ディスク構成要素をミラーディスク装置とすることで、異なるポートの複数ディスク装置の同時故障が起きても瞬時に上位装置の要求に対応することができ、極めて高い信頼性を得ることができる。

【0120】またディスク装置のエラー回数が閾値に達したときに将来故障する可能性が高いものと予測して、故障発生に対処できるようにミラーディスク装置における現用と予備の切替を予め行っておくことで、実際にディスク故障となってもディスク装置の機能を失うことなく瞬時に上位装置の要求に対し対応することができる。

【0121】更にディスクモジュールを2つ実装可能なディスクユニットを用いてディスクアレイ筐体を構成することでディスクアレイの実現が簡単にでき、またディスクアレイの増設や構成変更も簡単にできる。更にまた、構成要素をミラーディスク装置としたことに加えてパリティデータも併せて格納することで、ミラード化されたRAID3あるいはRAID5の動作形態を実現して極めて高い信頼性を得ることができる。

【0122】更に、ミラード化されない通常のディスク装置を用いたディスクアレイについて、ランクごとに少なくとも1台の予備用ディスク装置を設け、特定のディスク装置で予測障害が判定されたときに予備用のディスク装置を選択してデータを複写し、複写後に障害判定対象となったディスク装置と組み合わせて動的にミラーディスク装置を構成することで、通常のRAID構成における予備用ディスク装置を含めたディスク台数と同等の台数で、実質的に全構成要素をミラーディスク装置とした場合と同等の信頼性を低コストで実現することができる。

#### 【図面の簡単な説明】

【図1】本発明の原理説明図

【図2】本発明のハードウェア構成を示した実施例構成図

【図3】図2のミラード制御部の実施例構成図

【図4】本発明のディスクアレイ筐体の実施例構成図

【図5】図4の筐体の実装するディスクユニットを取出して示した説明図

【図6】本発明の動作機能を示した説明図

【図7】本発明で用いるデバイス管理テーブルの説明図

【図8】ディスク障害を予測判定した場合のデバイス管理テーブルの説明図

【図9】本発明におけるストライピングデータ格納状態の説明図

【図10】本発明の処理動作を示したフローチャート

【図11】本発明のデバイス検査処理を示したフローチャート

【図12】本発明の復旧処理を示したフローチャート

【図13】ミラーディスク構成でパリティを格納する本発明の実施例構成図

【図14】図13におけるRAID3の動作形態の説明図

【図15】図13におけるRAID5の動作形態の説明図

【図16】ミラード化しないディスクアレイ装置の構成変更を示した実施例構成図

【図17】ポート構成によりミラード化した本発明の実施例構成図

【図18】図17におけるデータストライピングの説明図

【図19】動的なミラーディスクを構成制御する本発明の他の実施例構成図

【図20】図19の動的なミラーディスク構成制御を示したフローチャート

#### 【符号の説明】

10：ホストコンピュータ（上位装置）

12：コントローラ（制御手段）

14：チャネル装置

16：チャネルインタフェースバス

18：ディスクアレイ

20：MPU

22：チャネルインタフェース制御部

24：制御記憶部

26：データバッファ

28：デバイスインタフェース制御

30-1～30-5：ミラード制御部

32-1～32-38：ディスク装置

34：内部バス

36-1～36-34：ミラーディスク装置

38：切替制御部

40-1～40-5：ディスク制御部

42-1, 42-2：デバイスバス

44：ディスクアレイ筐体

46-1, 46-10：ディスクユニット

48-1～48-5：フィルタ

50-1, 50-2：3.5インチ・ディスクモジュール

52-1, 52-2：モジュール回路基板

54：モジュール実装基板

56：電源ユニット

58：冷却ファン装置

60：ディスクアレイ制御部

62：デバイス管理テーブル

64：アクセス判定部

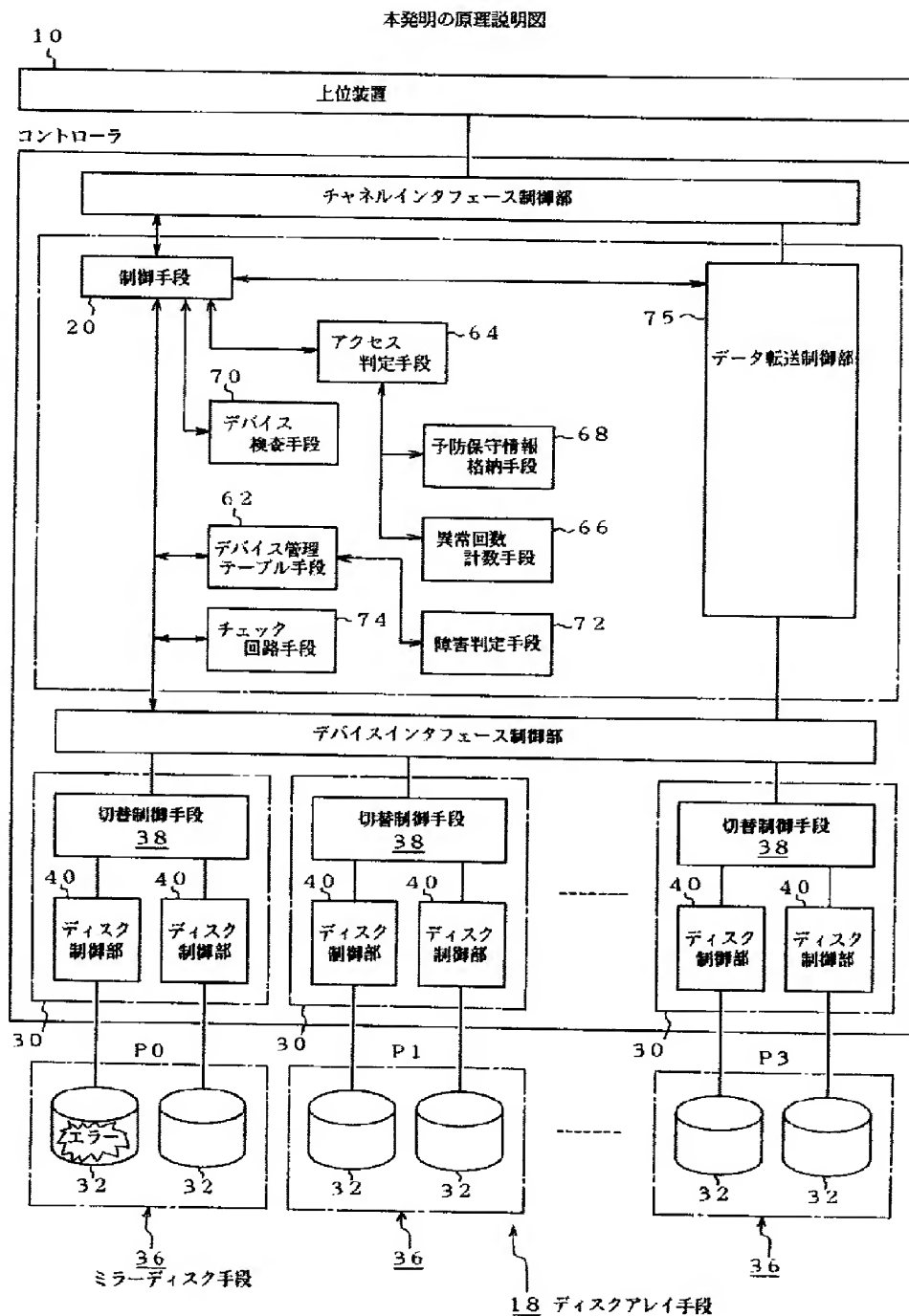
66：異常回数計数部

68：予防保守情報格納部

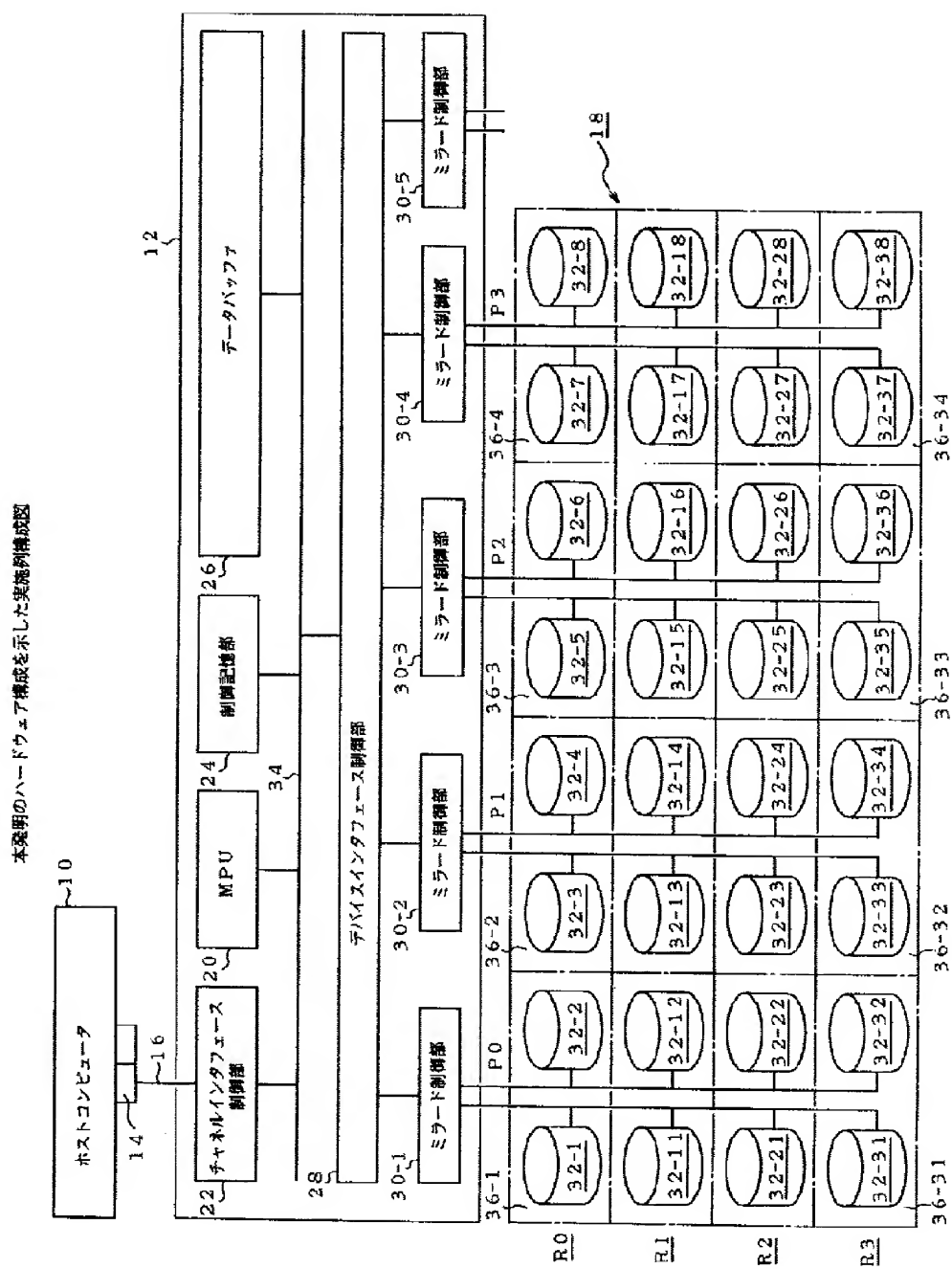
70：デバイス検査部

76: 論理ブロック (ホストブロック)  
78: ストライプデータ  
80: ダブルストライプデータ

【図 1】

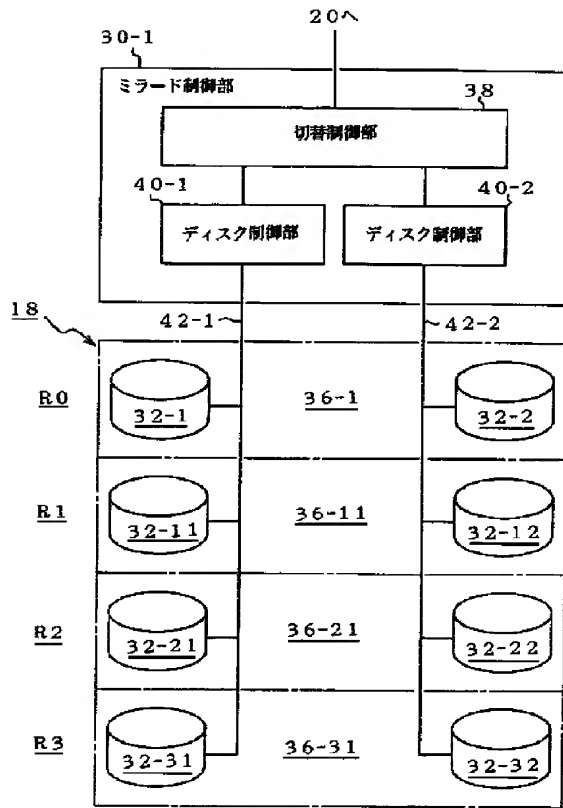


【図 2】



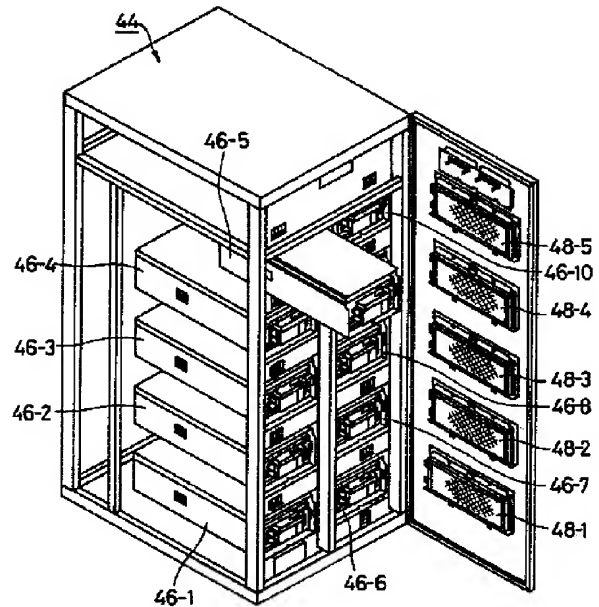
【図 3】

図 2 のミラード制御部の実施例構成図



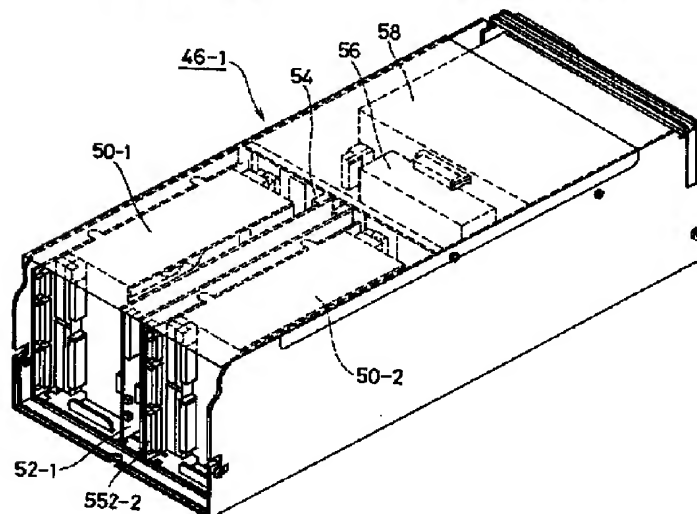
【図 4】

本発明のディスプレイ筐体の実施例構成図



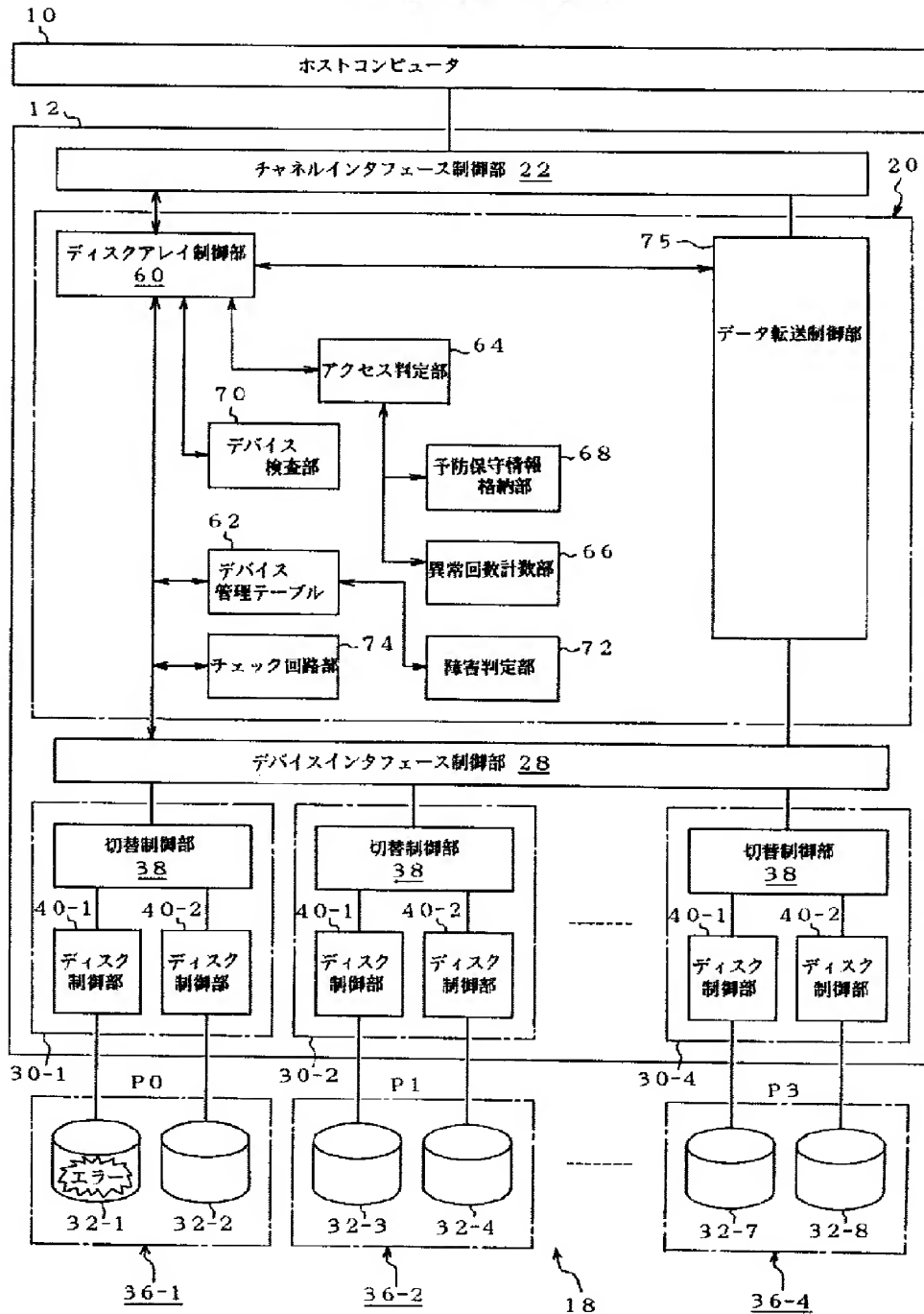
【図 5】

図 4 の筐体へ突装するディスクユニットを取出して示した説明図



【図6】

本発明の動作機能を示した説明図



【図 7】

本発明で用いるデバイス管理テーブルの説明図

62

ミラード ID	デバイス ID	ランク	ポート	現用フラグ	予備用フラグ	異常回数	故障フラグ
00	00	00	00	1	0	0	0
	01			0	1	0	0
01	02		01	1	0	0	0
	03			0	1	0	0
02	04		02	1	0		0
	05			0	1	0	0
03	06		03	1	0	0	0
	07			0	1	0	0
04	08	02	00	1	0	0	0
	09			0	1	0	0
05	10		01	1	0	0	0
	11			0	1	0	0
06	12		02	1	0	0	0
	13			0	1	0	0

【図 8】

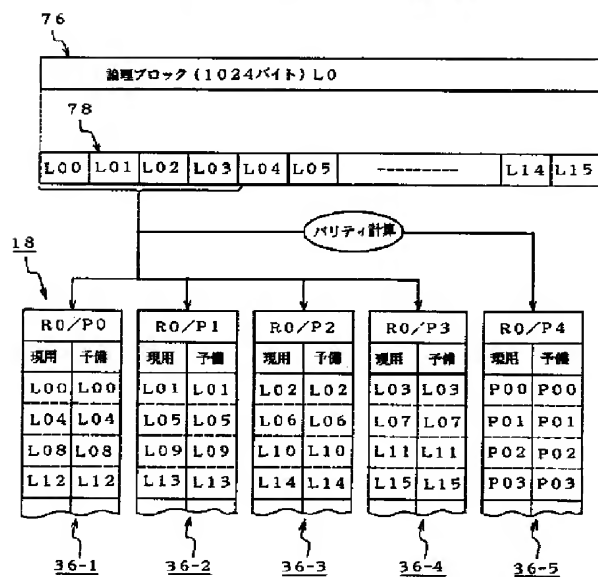
ディスク障害を予測判定した場合のデバイス管理テーブルの説明図

62

ミラード ID	デバイス ID	ランク	ポート	現用フラグ	予備用フラグ	異常回数	故障フラグ
00	00	00	00	0	0	10	1
	01			1	0	0	0
01	02		01	1	0	0	0
	03			0	1	0	0
02	04		02	1	0	1	0
	05			0	1	0	0
03	06		03	1	0	0	0
	07			0	1	0	0
04	08	02	00	1	0	0	0
	09			0	1	0	0
05	10		01	1	0	2	0
	11			0	1	0	0
06	12		02	1	0	0	0
	13			0	1	0	0

【図 14】

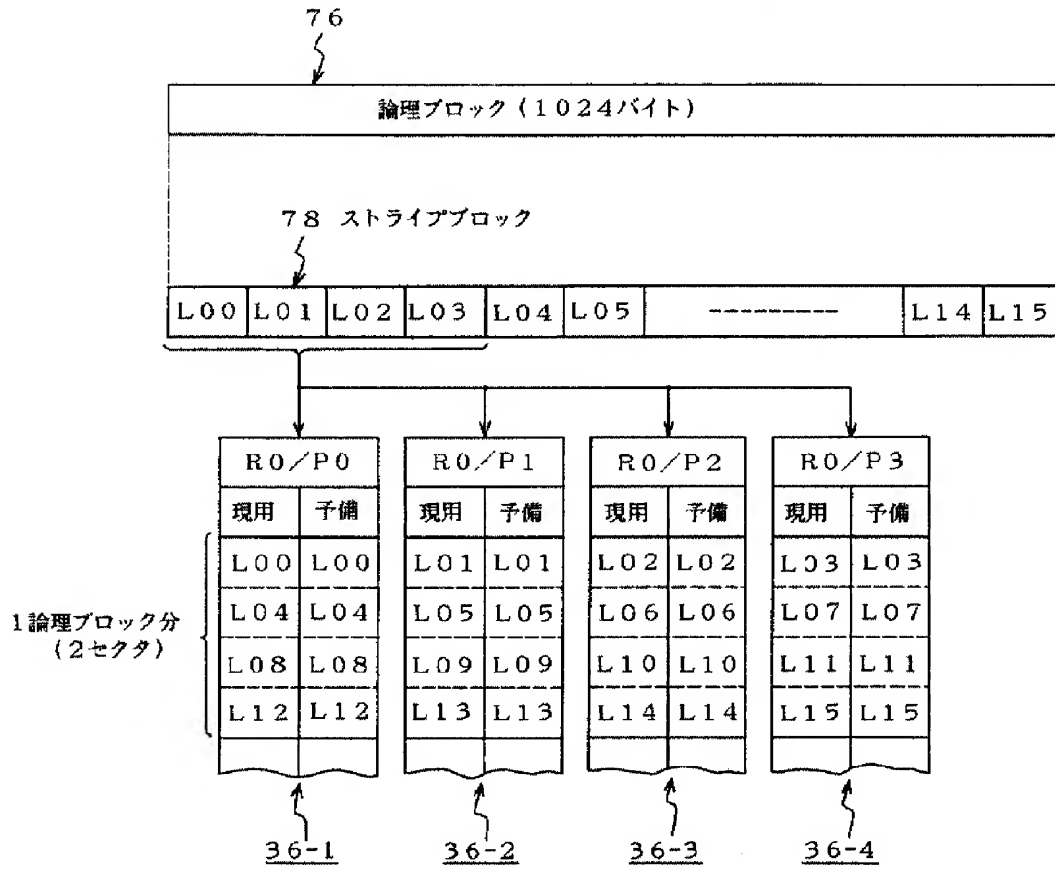
図 13 における RAID3 の動作形態の説明図





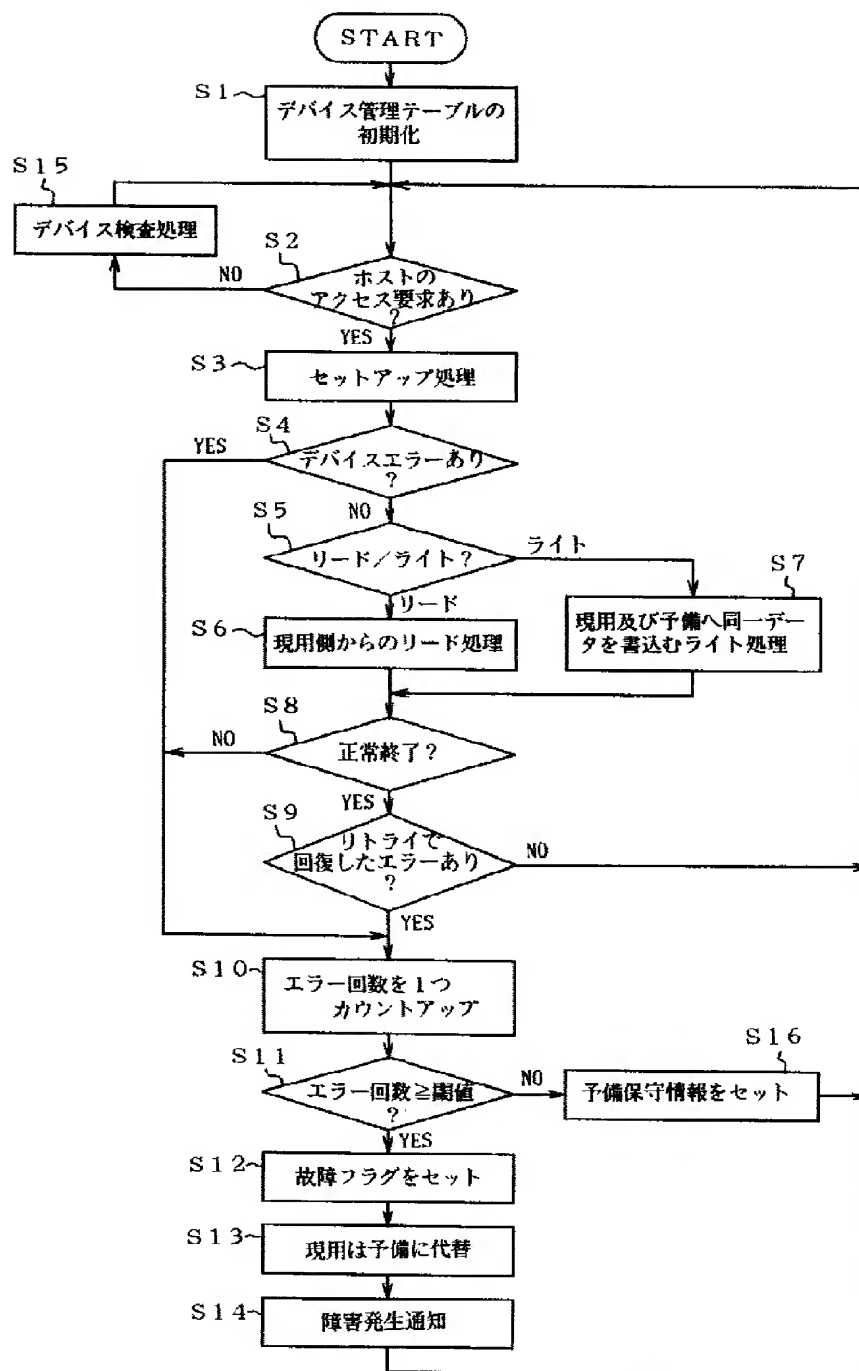
【図9】

本発明におけるストライピングデータ格納状態の説明図



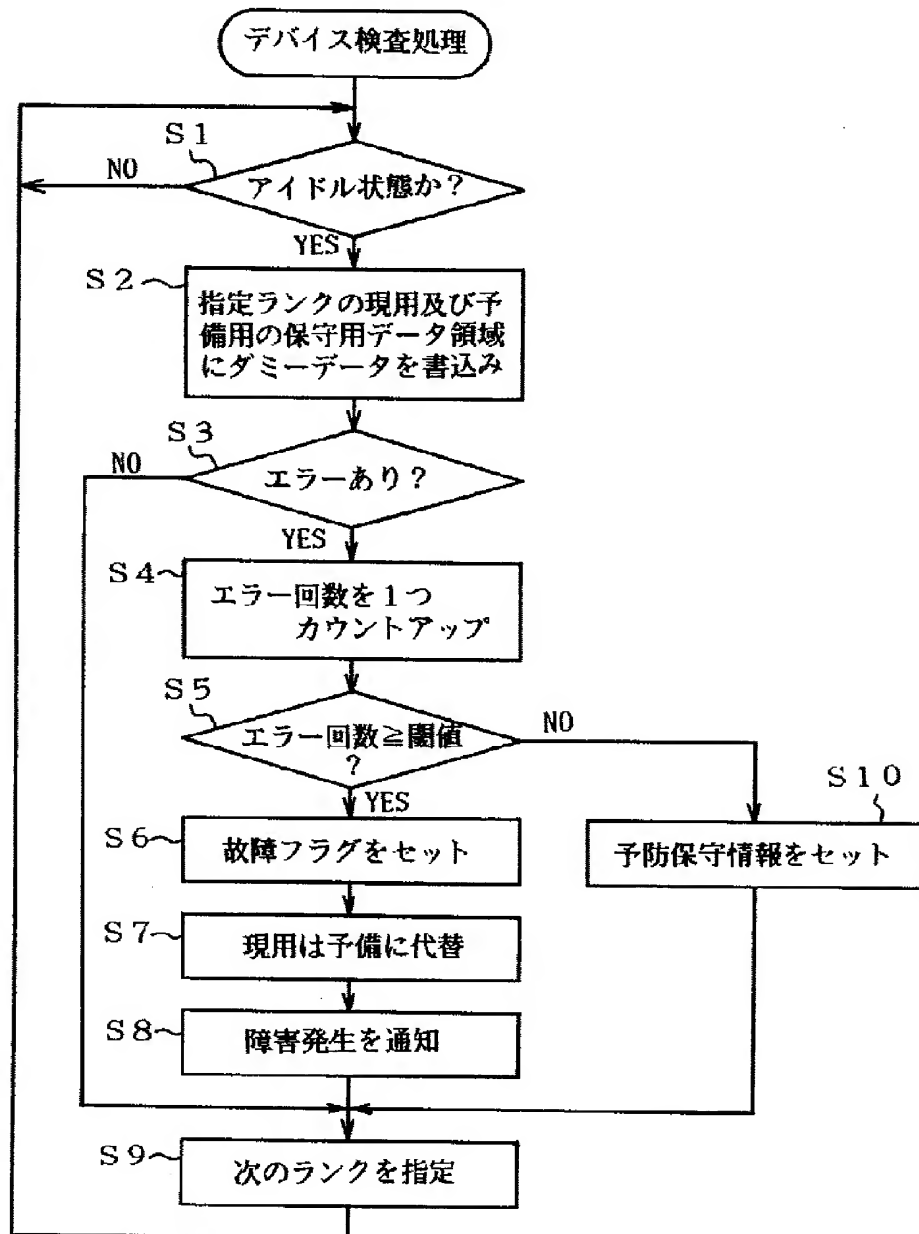
【図10】

本発明の処理動作を示したフローチャート



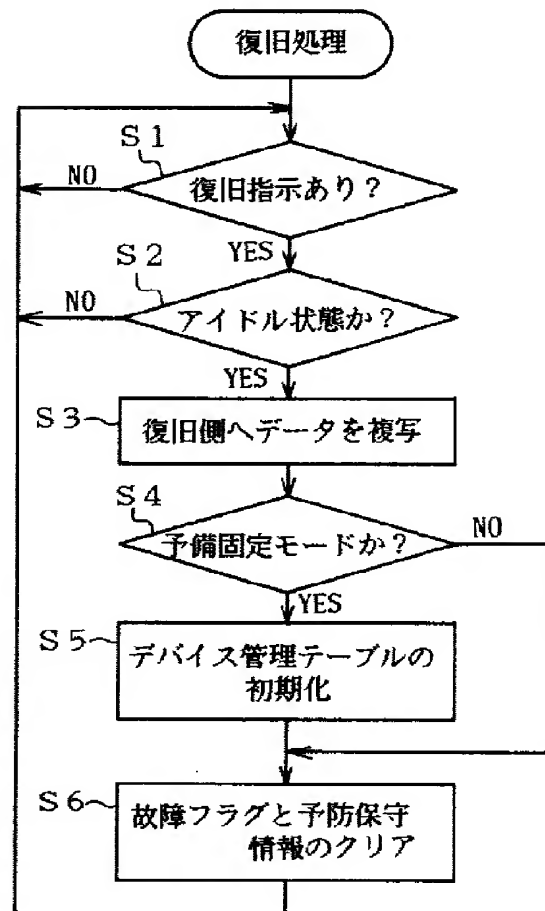
【図11】

本発明のデバイス検査処理を示したフローチャート

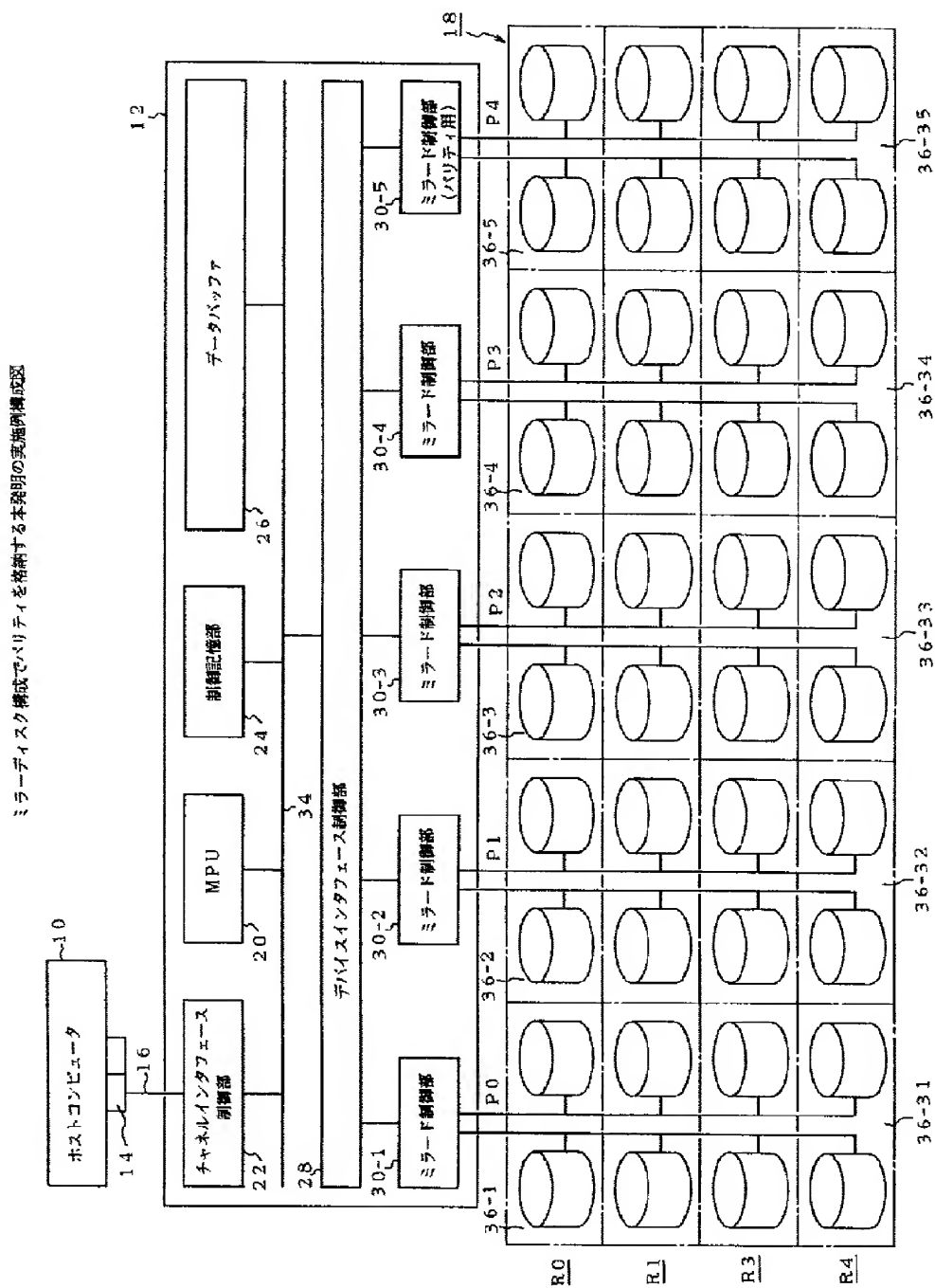


【図 12】

本発明の復旧処理を示したフローチャート

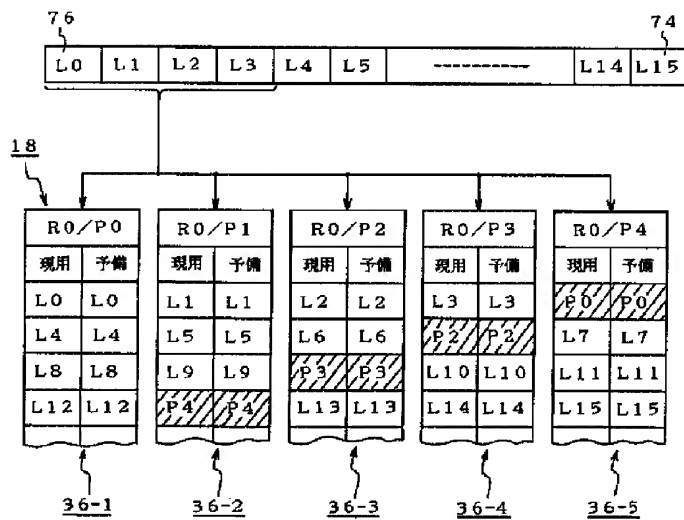


【図13】



【図15】

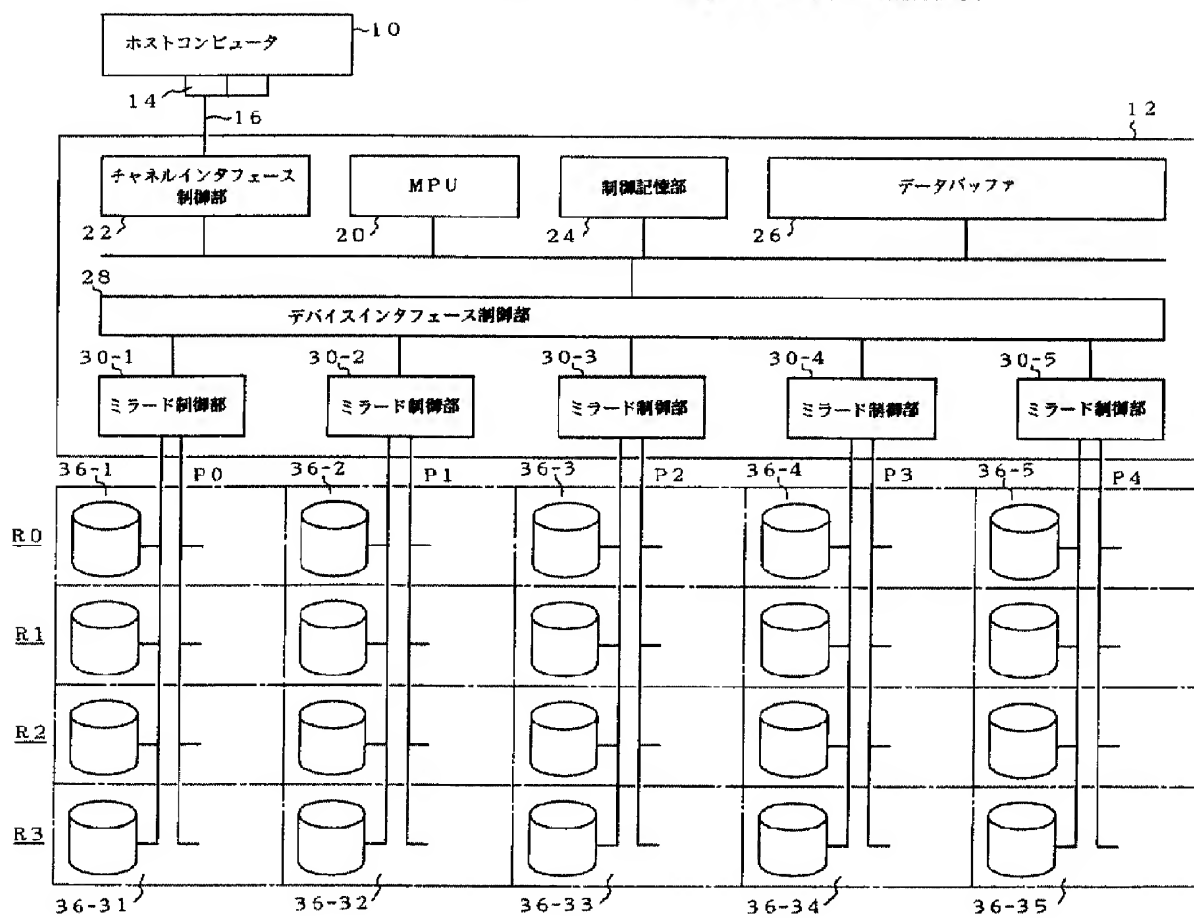
図13におけるRAID5の動作形態の説明図



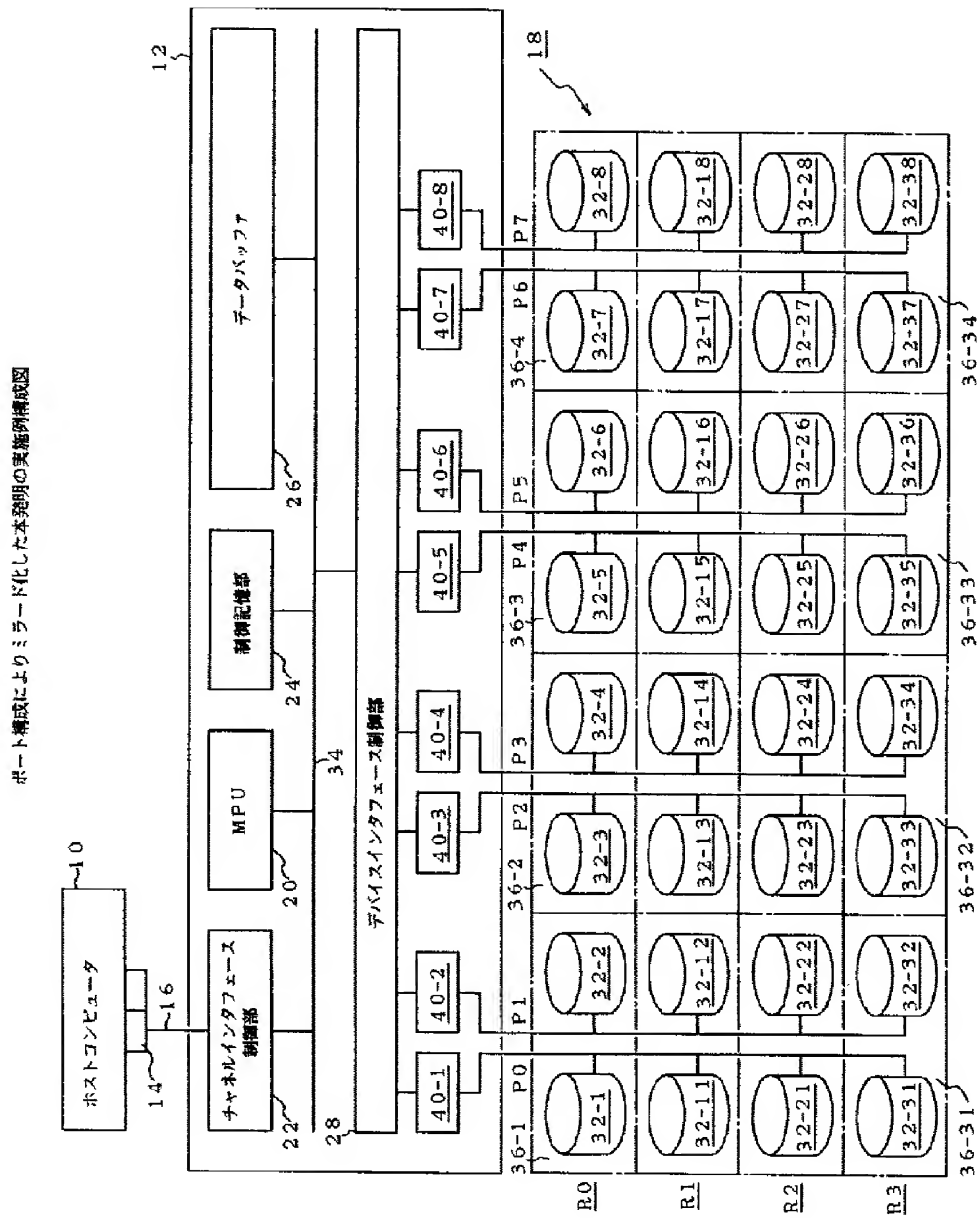


【図16】

ミラード化しないディスクアレイ装置の構成変更を示した実施例構成図

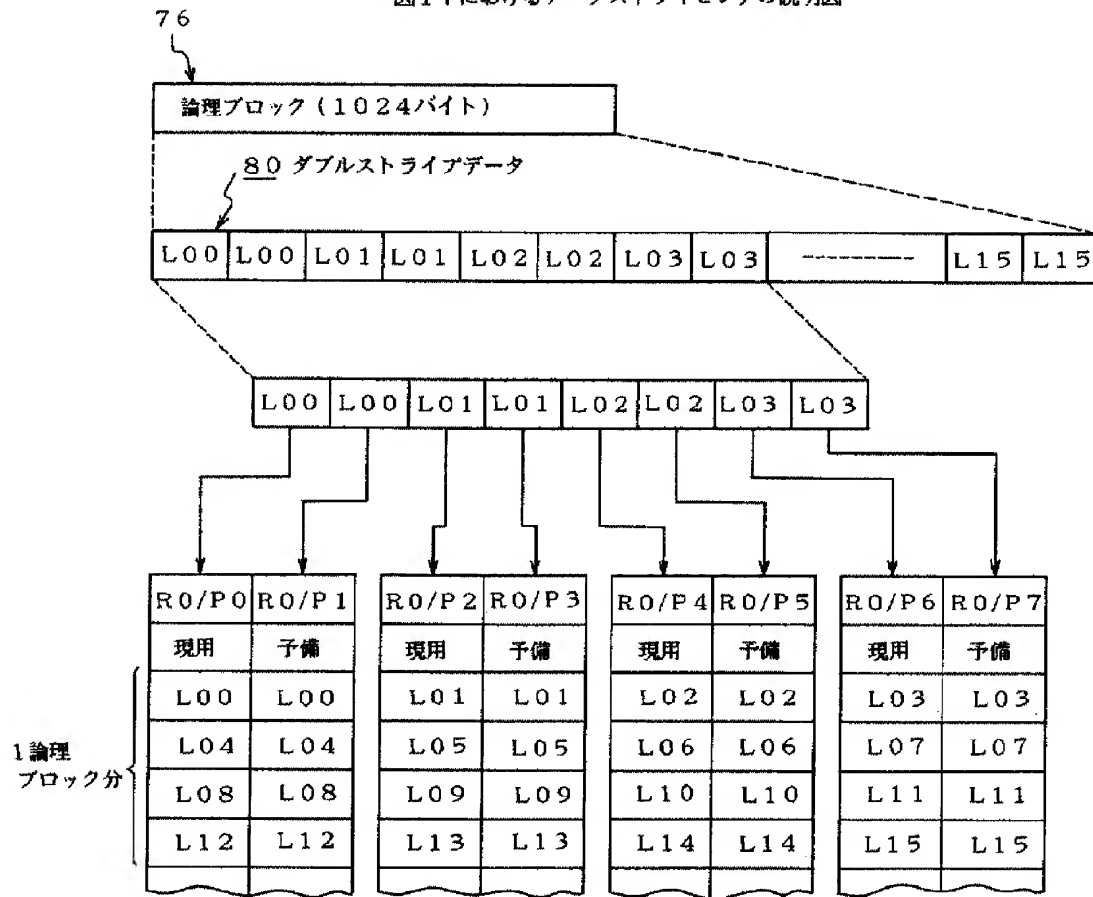


【図17】



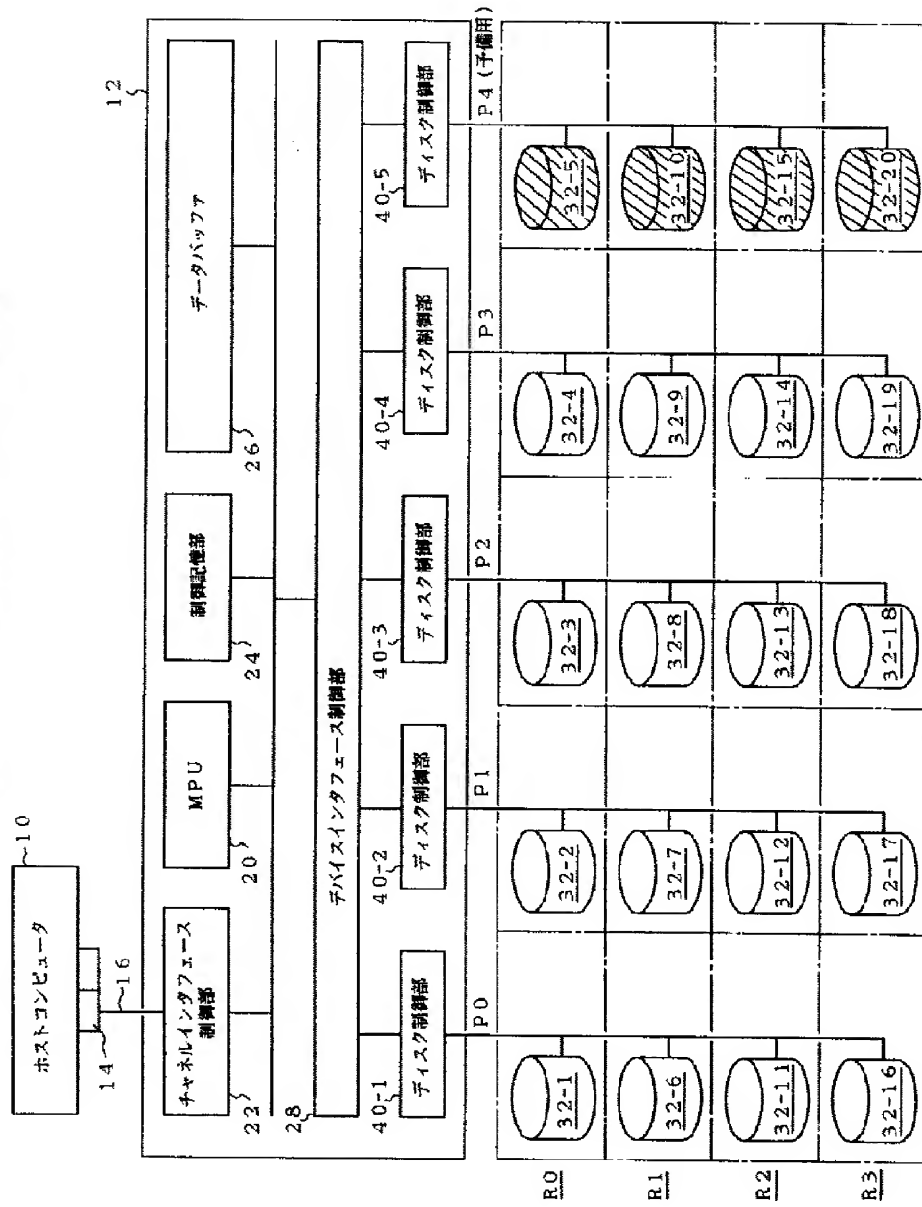
【図18】

図17におけるデータストライピングの説明図



【図19】

動的なミラーディスクを構成制御する本発明の他の実施例構成図



【図 20】

図 19 の動的なミラーディスク構成制御を示したフローチャート

